UNIVERSITÉ DE GRENOBLE

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Signal, Image, Parole, Telecom (SIPT)**

Arrêté ministériel : 7 août 2006

Présentée par

Amélie LELONG

Thèse dirigée par **Gérard BAILLY**

préparée au sein du **Département Parole & Cognition (DPC) de GIPSA-Lab**
dans **l'École Doctorale Electronique, Electrotechnique, Automatique & Traitement du Signal (EEATS)**
financée par **une allocation doctorale de recherche de la Région Rhône Alpes**

Convergence phonétique en interaction

Phonetic Convergence in interaction

Thèse soutenue publiquement le **3 juillet 2012**
devant le jury composé de :

M Gérard BAILLY
Directeur de recherche, GIPSA-Lab, Grenoble, (Directeur de thèse)
Mme Laurence NIGAY
Professeur, LIG, Grenoble, (Présidente)
M Jean-François BONASTRE
Professeur, LIA, Avignon, (Rapporteur)
M Noël NGUYEN
Professeur, LPL, Aix-en-Provence, (Rapporteur)
Mme Martine ADDA-DECKER
Directeur de recherche, LPP, Paris, (Examinateur)

Résumé

Giles a introduit en 1987 la théorie de l'accommodation communicative selon laquelle des locuteurs en interaction ont tendance soit à rapprocher leurs caractéristiques, vocales ou autres, de celles de leur partenaire, soit à accentuer la différence initiale. Ces stratégies d'adaptation auraient alors un but communicatif. Il semble donc judicieux de doter les agents conversationnels –animés ou robotisés– de ces mêmes stratégies. Nous avons donc mis en place un paradigme appelé « Dominos verbaux » pour caractériser une de ces stratégies d'adaptation communément appelée convergence phonétique (i.e. rapprochement des caractéristiques vocales). Le principe du jeu est simple : les sujets doivent choisir entre deux mots celui qui commence par la syllabe finale du mot précédemment énoncé par leur interlocuteur. Ce paradigme nous permet alors de contrôler le nombre d'exemplaires de chaque phonème collectés pour caractériser le phénomène. Il nous permet également de varier les conditions d'enregistrements pour obtenir des espaces phonétiques de référence (pendant le pré-test), étudier le phénomène en interaction (en face-à-face ou en interaction ambiante) ou pour étudier le phénomène d' « after-effect » pendant un post-test. Nous avons, dans un premier temps, étudié le phénomène de convergence phonologique pour ensuite nous concentrer sur la convergence phonétique. Nous avons développé deux méthodes objectives de caractériser la convergence phonétique. La première méthode, qui demande une segmentation précise des signaux a priori, est basée sur une analyse discriminante linéaire entre les coefficients MFCC des voyelles extraits des pré-tests des partenaires de chaque paire. Cette analyse nous permet de calculer un taux de convergence correspondant à un rapport de distance spectrale pendant l'interaction et pendant les pré-tests. La deuxième méthode, qui ne nécessite pas de traitement a priori, utilise la reconnaissance du locuteur. Le taux de convergence est calculé en fonction des rapports de log-vraisemblance obtenus en procédant à de la reconnaissance de locuteur croisée. Les deux méthodes nous montrent qu'il y a effectivement une adaptation des sujets pendant l'interaction. De plus, les taux de convergence obtenus avec chaque méthode sont corrélés. Nous pourrons donc utiliser la deuxième méthode pour caractériser la convergence phonétique en situation moins contrôlées. Les résultats obtenus démontrent que l'amplitude de la convergence est dépendante du sexe des paires (on obtient les meilleurs résultats pour des paires de femmes) et de la distance sociale (les taux de convergence pour des inconnus sont inférieurs à ceux entre des amis eux-mêmes inférieurs aux taux obtenus pour des personnes d'une même famille). Nous avons également étudié le lien entre l'amplitude de la convergence et la fréquence lexicale ainsi qu'avec le temps. Les résultats ne montrent cependant pas de tendance claire et méritent d'être approfondis. Enfin, nous avons étudié l'impact de la connaissance de la cible linguistique sur la production en demandant aux sujets de répéter le domino précédemment énoncé par leur partenaire avant de prononcer le leur. Les résultats ne sont cependant pas concluants. Dans une dernière partie, nous avons voulu caractériser la convergence phonétique de manière subjective. Nous avons donc mis en place plusieurs tests AXB mais les résultats obtenus ont démontré que ce type de tests n'était pas adapté à notre problématique. Nous avons donc développé un nouveau test de perception basé sur la détection « en ligne » d'un changement de locuteur. En effet, si les sujets se sont adaptés pendant l'interaction, il sera alors plus difficile de distinguer une transition entre les signaux provenant de l'interaction. Les résultats obtenus confirment cet effet, les sujets sont donc sensibles à la convergence phonétique. Le même type de test a été mis en place avec des signaux de synthèse adaptative, créés à partir de la modélisation HNM. Nous avons obtenus des résultats comparables démontrant ainsi la qualité de notre synthèse adaptative.

Mots Clés

Convergence phonétique, accommodation, adaptation mutuelle, alignement, reconnaissance du locuteur, analyse discriminante linéaire, convergence phonologique, reconnaissance de parole, dominos verbaux, synthèse adaptative, modélisation harmonique plus bruit, perception, détection de changement de locuteur.

Abstract

The communicative accommodation theory has been introduced by Giles in 1987. It postulates that speakers, in interaction, will tend to move their –vocal or other– features closer to those of their partner or to accentuate their differences. These adaptation strategies would have a communicative goal. So it seems appropriate to equip the conversational agents –virtual or robotic– with these strategies. So, we have developed a paradigm called “Verbal Dominoes” to characterize one of these strategies named phonetic convergence. The rule of the game is quite simple: speakers have to choose between two words the one that begins with the same syllable as the final syllable of the word previously uttered by the interlocutor. This paradigm permitted us to control the number of exemplars of each phoneme to characterize the. Moreover, it allows us to use different recording types to get the references for each speaker (during the pre-test), to study the phenomenon in interaction (in face-to-face or in ambient perturbation) or to study the “after-effect” phenomenon during a post-test. First, we have studied the phonological convergence (i.e. cross-categorical like different accents) in order to focus on the phonetic convergence for the rest of the study. Two objective methods to characterize phonetic convergence have been developed. The first one, that requires an a priori segmentation and labeling of phonemes, uses a linear discriminant analysis. It is performed on the target MFCC parameters for each vowel to categorize each interlocutor’s vocalic space into two distinct groups. A normalized convergence rate is then computed by dividing the distance between targets produced during the interaction with that produced during the pre-test for the same word. The second method that does not require any a priori processing is based on speaker recognition techniques. Convergence rates are then taken as the relative quotient between differences of log-likelihood ratio obtained during pre-test and interaction by using cross recognition. Both methods show us that adaptation occurs in interaction. Furthermore, convergence rates obtained with the two methods are correlated. So we will be able to use the second one to characterize phonetic convergence in less well controlled conditions. Results show that phonetic convergence is stronger for pairs of the same sex (particularly for women) and depends on social distance (convergence rates increase while social distance decreases). The link between convergence rate and lexical frequency has been studied too. Its link with time has been also investigated. Nevertheless, there is no clear tendency in the results. They should be studied in detail. Then, we have asked our subjects to repeat domino that had just been uttered by their partner before pronouncing theirs. This new condition has been added to study the knowledge of the linguistic target impact but the results were inclusive. In a last part, we wanted to characterize phonetic convergence in a subjective way. We have developed some AXB tests but we have concluded that this type of test was not suitable for our problem. So, we have used a novel perceptual test that we have called *speaker switching.* The listeners’ task is simply to press a key each time they perceive/suspect a speaker switch In fact, if listeners are sensitive to phonetic convergence, it should be more difficult for them to detect a transition between signals from interaction rather than those from pre-tests. Results confirm this effect. The same kind of test has been designed with synthetic data (obtained with the Harmonic plus noise model) that simulate a symmetrical 20% convergence between speakers. Results are comparable confirming the good quality of our adaptive synthesis.

Keywords

Phonetic convergence, accommodation, mutual adaptation, alignment, speaker recognition, linear discriminant analysis, phonologic convergence, speech recognition, verbal dominoes, adaptive synthesis, Harmonic plus noise model, perception, speaker switching.

Remerciements

Je tiens d'abord à remercier la région Rhône-Alpes qui a financé cette thèse grâce à une allocation doctorale de recherche.

Je voudrais ensuite remercier chaleureusement Gérard Bailly de m'avoir donné la chance de travailler sur ce sujet novateur et passionnant et également pour son soutien au cours de ces trois années de thèse.

Je tiens également remercier tous les membres du jury, Laurence Nigay en tant que présidente de mon jury de thèse, Jean-François Bonastre et Noël Nguyen, qui ont accepté de rapporter sur ce manuscrit et Martine Adda-Decker d'avoir participé en tant qu'examinatrice.

Je voudrais aussi remercier le GIPSA Lab et le département Parole et Cognition qui a financé une prolongation de deux mois supplémentaires afin que je puisse terminer la rédaction de mon manuscrit dans de meilleures conditions.

Je tiens de manière générale à remercier tous les membres du laboratoire GIPSA, et en particulier ceux du Département Parole et Cognition, que j'ai eu la chance de croiser. J'adresse une pensée particulière à tous mes amis doctorants, surtout à Audrey, Benjamin et Olha qui ont su m'écouter et me soutenir lorsque j'en avais besoin, ainsi qu'à mes amis de la pause café, Yo, Nico, Xavier, Louis, Guillaume et Rogelio. Je remercie également Maëva Garnier, Frédéric Elisei, Thomas Hueber et Sascha Fagel pour les conversations et les conseils qu'ils ont pu me prodiguer.

Je voudrais également adresser un remerciement à tous mes collègues des RJCP avec qui j'ai vécu une aventure sympathique et enrichissante: Atef, Benjamin, Hien, Mathilde, Rosario, Sandra (et une dédicace spéciale pour Louis).

Je tiens aussi à remercier tous mes amis qui se sont impliqués à leur manière mais également concrètement dans ma thèse, Laure, Noémie, Elsa, Laetitia, Camille, Mathilde, Fanny, Mickaël, Fabien, Johan et beaucoup d'autres. J'en profite pour remercier toutes les personnes qui ont participées aux diverses expériences que j'ai menées.

Enfin, je voudrais remercier chaleureusement ma deuxième famille, Isabelle, Michel, Perrine, Aurélien et Pauline, qui m'entoure et m'encourage au quotidien. Maintenant, je voudrais dédier ce travail à ma famille, que je n'oublie pas même s'ils sont loin, et plus particulièrement à mes parents, qui m'ont toujours soutenu et m'ont accompagnée tout au long de mes études. Ils m'ont toujours encouragée et m'ont poussée à donner le meilleur de moi-même. Un grand merci également à Magali, Ludo, Aline et Greg.

Le mot de la fin sera pour Sylvain. Comme tu n'aimes pas les grandes formules, je vais essayer de faire simple. MERCI d'être là pour moi tous les jours, pour ton implication dans mon sujet de thèse, les conversations que nous avons pu avoir même si apparemment tu n'y comprenais pas grand chose =). Tu as toujours su trouver les mots justes pour me remonter le moral et me donner l'envie d'avancer. Ce travail n'aurait pas été le même sans ton soutien et ton amour, merci et l'aventure continue…

Table des matières

Tables des Figures

Tables des Tableaux

Acronymes

AMORCES: Algorithmes et MOdèles pour un Robot Collaboratif Eloquent et Social
ANOVA : ANalysis Of Variance
CAT : Communication Accommodation Theory
DCT: Discrete Cosine Transform
DTW: Dynamic Time Warping
EM: Expectation-Maximization
f0: Fréquence fondamentale
F1: Premier formant
F2: Deuxième formant
F3: Troisième formant
GMM : Gaussian Mixture Model ou Modèle à mélange gaussien en français
GMUP: Group 'em up
HMM: Hidden Markov Modèle ou Modèle de Markov caché en français
HNM: Harmonic plus Noise Model
HTK : Hidden Markov Model Toolkit
IAT: Implicit Association Task
ISI : Inter Stimuli Interval
LDA : Linear Discriminant Analysis (Analyse Discriminante Linéaire)
LLR : Log Likelihood Ratio
MFCC: Mel-Frequency Cepstral Coefficient
SAT : Speech Accommodation Theory
SPL : Sound Pressure Level (Niveau de pression acoustique)
TTT: Turn Taking Time
VCV: Voyelle Consonne Voyelle
VOT: Voice Onset Time

Introduction

Les systèmes d'interactions homme-machine sollicitent et mobilisent de plus en plus de modalités sensorielles et motrices. Grâce à des technologies de capture de mouvement pervasives – capturer les mouvements du corps avec les Wii, Kinect et autres dispositifs d'accélérométrie ou de télémétrie infrarouge ou des yeux avec les oculomètres embarqués – ou au corps – capture et exploitation de données physiologiques ou des activités cérébrales – les systèmes d'information – jeux ludiques ou « sérieux » – peuvent percevoir de manière plus en plus fine les activités des humains présents dans leur environnement, exploiter leurs actions et leurs réactions aux actions engagées par ces systèmes. La robustesse des boucles de perception-action, la pertinence du liage entre signaux émis par le monde physique et ceux synthétisés par les systèmes numériques conditionnent la pertinence et l'efficacité des services offerts.

Ces boucles de perception-action sont complexes : la planification et l'exécution d'actions chez l'humain est le résultat tangible d'un vaste ensemble de processus mentaux impliquant perception, motricité, attention, mémoires (long-terme, court-terme, de travail, épisodique), apprentissage, raisonnement, émotion, langage, pensée abstraite, volition, etc. Les plans d'action sont sélectionnés et paramétrés en fonction de caractéristiques physiques de l'environnement (taille, distance des objets) aussi bien que de contraintes plus cognitives telles qu'imposées par les usages linguistiques, sociaux ou culturels.

L'apprentissage de ces processus par chaque individu est largement implicite. Il est le résultat d'un héritage génétique, des expériences et initiatives – récompensées ou non par l'environnement ou le milieu social – personnelles ainsi que du conditionnement social. Il est donc assez tentant pour les systèmes interactifs de s'appuyer sur les acquis de cette pratique routinière de gestion des boucles de perception-action entre humains, ne nécessitant pas ou peu d'adaptation, pour interagir avec des partenaires humains. Ainsi, les ressources cognitives mobilisées par les partenaires pour s'approprier les nouveaux outils et services seront limitées, l'usage intuitif (Thórisson, 2002)et l'adaptation rapide.

Les systèmes bio-inspirés se nourrissent ainsi de données psychophysiques et comportementales collectées sur l'humain engagé dans des interactions réelles ou simulées avec les objets et les agents présents dans son environnement cyber-physique proche. Notre travail s'inscrit dans la perspective de doter un agent conversationnel – robotique ou virtuel – de la capacité à converser de manière intuitive et fluide avec de multiples partenaires humains en s'adaptant à la tâche, la situation d'interaction et l'environnement dans lesquels sont conduit cette interaction ainsi qu'aux capacités sensori-motrices et compétences cognitives de son interlocuteur et en donnant les gages sensibles de cette attention et conscience de l'autre.

Le travail présenté dans cette thèse se concentre sur l'étude d'un phénomène d'adaptation qui fait l'objet d'un regain important de recherches depuis quelques années : l'adaptation phonétique ou le rapprochement de caractéristiques vocales des locuteurs en interaction.

Après une revue de l'état de l'art, où nous essaierons de proposer une taxonomie des situations d'interaction homme-homme et homme-machine étudiées jusqu'ici et les propriétés émergentes du

couplage verbal, le manuscrit comporte trois chapitres : le chapitre deux décrit les différents types de scénarios utilisés pour récupérer des corpus permettant l'étude de la convergence, le chapitre trois montre les méthodes basées sur des mesures acoustiques utilisées pour caractériser l'adaptation en interaction et le chapitre quatre présente quelques tests de perception . Nous conclurons enfin sur le fonctionnement et l'apport de la convergence d'après nos résultats et développerons les perspectives qui découlent de ces travaux.

Chapitre 1 Etat de l'art

La notion de convergence est utilisée dans de nombreux domaines scientifiques, que ce soit en mathématiques, économie, biologie, géologie voire géopolitique. Elle désigne de manière générale le rapprochement, la diminution de distance entre les caractéristiques quantitatives ou propriétés plus qualitatives d'éléments en interaction, ceci de manière systémique ou récurrente. La convergence phonétique désigne le rapprochement de caractéristiques vocales – patrons de respiration, caractéristiques spectrales des sons, indices temporels, traits distinctifs, prosodie, etc. – de locuteurs interagissant entre eux ou avec des dispositifs technologiques vocaux.

Il est bien difficile de dissocier les termes d'un possible emprunt à l'autre – cet effet « magnet » d'attirance vers le comportement d'autrui – des termes d'adaptation conjointe à une situation donnée. Sommes-nous attirés vers l'autre dans une situation et pour un but donné ou sommes-nous tous attirés vers des comportements plus collectifs – donc plus similaires – de manière à répondre à une situation commune ? On essaiera de distinguer dans la suite de la présentation ce qui relève de l'adaptation en ligne – spécifique à l'interlocuteur et à ses caractéristiques – de ce qui est plus contingent à la situation de communication.

Avant de discuter plus avant les diverses expérimentations menées dans la littérature pour mettre en valeur les phénomènes de convergence spécifiques à la parole et leurs possibles motivations, nous allons présenter les diverses notions de mimétisme, imitation, alignement, synchronisation, etc. qui touchent notre domaine de recherches et montrer que le mécanisme d'alignement comportemental est assez général et qu'il a été mis en évidence pour de multiples dimensions.

1.1 Mimétisme et imitation

Le **mimétisme** (mimicry en anglais) est une stratégie adaptative d'imitation employé de manière très large par les espèces animales à toutes les échelles, y compris au niveau des cellules. Elle permet notamment à une espèce (a) d'échapper à d'éventuels prédateurs (on peut penser à l'utilisation de l'allocryptie par le Phrygane rhombifère, à la simulation de la thanatose par certains reptiles ou à la présence d'ocelle[1] dans de nombreuses espèces de poissons)

[1] Tâche arrondie qui sert de leurre ou de moyen d'intimidation sur la peau ou les ailes d'animaux simulant un œil, soit pour surprendre le prédateur éventuel en lui présentant un regard sans commune mesure avec le potentiel de la proie, soit en lui suggérant un autre emplacement de la tête.

(b) de masquer temporairement son identité pour mettre en confiance des proies (on peut penser à certains parasites qui reproduisent les antigènes de leur hôte)
(c) de bénéficier de soins d'autres espèces (les coucous sont ainsi connus pour pondre dans les nids des autres des œufs de même couleur que l'espèce hôte, leurs poussins sachant de plus imiter le cri des poussins de l'oiseau hôte qui les nourrira !)

Il est clair que le comportement similaire des individus confrontés à une même situation, induit par cette adaptation sélective à l'environnement, ne peut être en aucun cas assimilé à une imitation, qui est la copie d'une action d'autrui, inhabituelle ou nouvelle, sans recours à un comportement d'origine instinctive (Thorpe, 1967; et repris par Serkhane, 2005). En fonction de l'environnement, les sujets adaptent de manière assez similaire leur manière de parler – comme par exemple hausser la voix dans le bruit vs. chuchoter en milieu calme – sans que cela soit induit par la présence de congénères ni en réponse à leur comportement. On parlera plutôt dans ce cas de contagion (Bard and Russell, 1999) qui sollicite un comportement appris, déclenché par l'observation de ce même comportement chez autrui ou provoqué par des causes identiques.

Figure 1. 1. Photos d'un nouveau-né de deux-trois semaines imitant les expressions faciales d'un adulte (Meltzoff and Moore, 1977; Meltzoff and Moore, 1983)

Lors d'une imitation, le modèle à copier n'est donc pas obligatoirement à chercher au sein d'une autre espèce. Il peut aussi être intra-spécifique, c'est-à-dire s'opérer au sein d'une même espèce, aussi bien pour rapprocher les sexes pendant la période de reproduction, le maintien de la cohésion au sein d'un groupe, hiérarchie et territorialité, liens parentaux, coopération, etc. : de nombreuses espèces utilisent l'imitation – notamment vocale (Baker, 1993) mais pas exclusivement – à des fins de communication intra-spécifique. La reproduction du comportement d'autrui peut ainsi communiquer la reconnaissance d'autrui comme faisant partie du groupe ou l'acceptation du partenaire. Ainsi si on peut supposer que le développement des chants particuliers aux espèces d'oiseaux est largement régulé par la maturation des organes (Podos, 1996) et les secrétions saisonnières d'hormones, des

variations géographiques, quasi-dialectales (Baptista and Petrinovitch, 1984; DeWolfe *et al.* ,1989; Putland *et al.* ,2006) sont observées.

1.2 Alignement et convergence

L'**alignement** désigne l'action de mettre en correspondance des séquences composées de plusieurs objets de manière à aligner les objets communs et faire ressortir des régions homologues ou similaires. En bioinformatique, l'alignement de séquences consiste par exemple à maximiser le nombre de coïncidences entre nucléotides ou acides aminés afin d'identifier des sites fonctionnels, de prédire la ou les fonctions d'une protéine. Pour ce qui est du comportement, l'alignement consiste à adopter la ligne de pensée, la manière d'agir d'un groupe ou d'une personne.

Une conversation, comme toute activité conjointe entre deux individus, suppose que les contributions de chacun sont liées, dépendantes des contributions de l'autre, par exemple en réponse à une question. Cette interdépendance ne réfère pas seulement au contenu mais aussi à la forme des contributions de chacun (Linell, 1998). Pickering & Garrod (2004; 2006) proposent ainsi que l'alignement linguistique soit la base d'une communication fructueuse. A la suite de Brennan & Clark (1996), les arguments expérimentaux des auteurs sont principalement ancrés sur les expressions référentielles portant sur des objets, impliqués dans les tâches collaboratives dans lesquelles ils impliquent leurs interlocuteurs. L'alignement peut cependant concerner des structures de plus haut niveau, notamment syntaxiques (Branigan *et al.* ,2000; Pickering *et al.* ,2000; Lockridge and Brennan, 2002). Il a également été observé au niveau des gestes iconiques, par exemple pour les gestes de la main accompagnant la parole (Mol, Krahmer and Swerts, 2009 ; Kimbara, 2008), le rire (Young and Frye, 1966), les expressions faciales (Bavelas, Lemery and Mullet, 1986) ou les émotions (Hatfield, Cacioppo and Rapson, 1994)

La **convergence** suppose une métrique objective ou subjective capable d'évaluer la distance entre objets (protéines, ontologies, sons, etc.) alignés. On distingue deux métriques :

- l'une temporelle, qui va évaluer le degré de coordination – on parle souvent de synchronisation – entre comportements ; Richardson et al (2008) ont ainsi étudié la synchronisation du balancement du centre de gravité de deux interlocuteurs suivant leur degré de visibilité mutuelle et d'interactivité.
- l'autre spatiale, qui va évaluer le degré de ressemblance entre les objets élémentaires comme par exemple pour les paramètres prosodiques.

1.3 Convergence phonétique

L'engouement pour l'étude de la convergence des réalisations phonétiques de locuteurs en interaction est massive mais assez récente. Elle a porté dans un premier temps essentiellement sur les paramètres prosodiques : volume (Kousidis et al. ,2008), registre de f0 (Gregory et al. ,1993; Gregory and Webster ,1996; Gregory Jr *et al.* ,2001), rythme (Matarazzo and Wiens, 1967; Street, 1984; Kousidis *et al.* ,2008; Edlund *et al.* ,2009), etc. Plus récemment, les caractéristiques segmentales (Delvaux and Soquet, 2007) des mots échangés voire des tours de parole ont été étudiées, ainsi que certaines caractéristiques phonatoires ou articulatoires comme le rythme respiratoire (McFarland, 2001) ou l'ouverture labiale (Gentilucci and Bernardis, 2007).

1.4 Mécanismes sous-jacents : automatisme vs. perméabilité cognitive

Plusieurs théories s'opposent pour expliquer le phénomène de convergence. D'un côté, il est considéré comme une réaction automatique de bas niveau, comme un effet de résonance à un stimulus, de l'autre il est défini comme une stratégie d'interaction volontaire. Bien qu'aucune étude ne puisse trancher de manière définitive entre mécanisme exogène versus endogène, les études ont montré que le phénomène de convergence restait complexe et qu'une théorie seule ne pouvait pas l'expliquer dans sa totalité. On peut alors se demander s'il n'y aurait pas en fait un couplage possible entre les théories existantes pour comprendre ce phénomène.

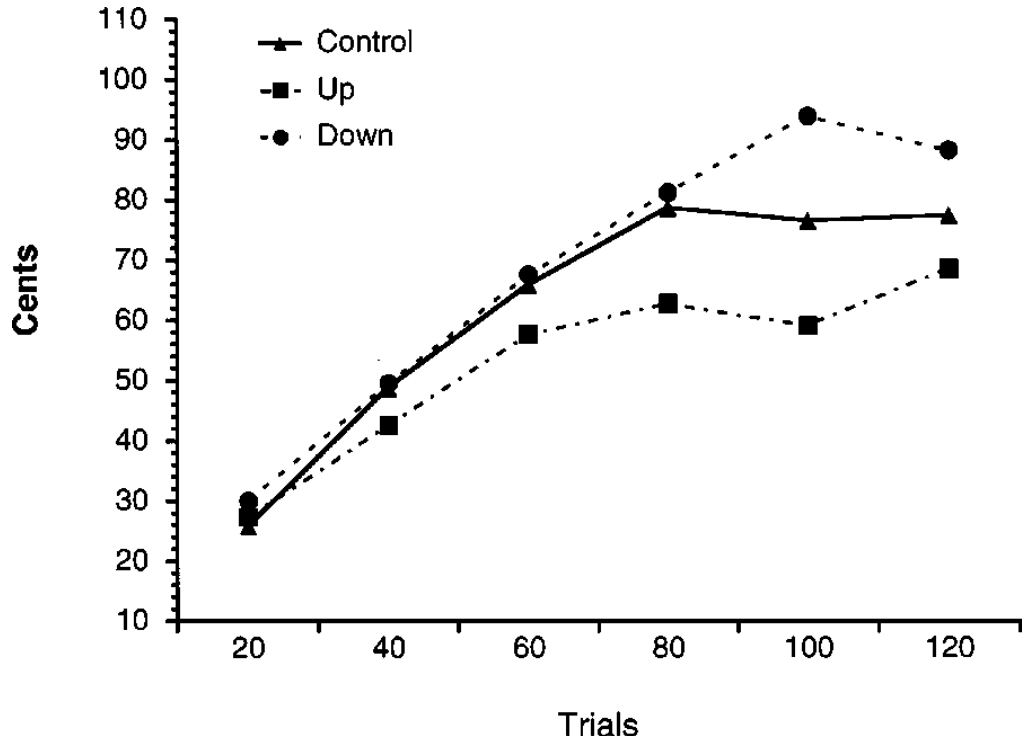


Figure 1. 2. Comparaison de f0 produites avec un retour auditif non perturbé (Control) versus celles produites avec un retour où le f0 est augmenté (Up) ou réduit (Down) de 1 cent par essai (Trial). La perturbation maximale de 100 cents est conservée pour les derniers 20 essais. Si les locuteurs ont une tendance générale à augmenter leur registre, ils ont aussi tendance à surcompenser la perturbation. (d'après Jones and Munhall, 2000).

1.4.1 Les perturbations du retour perceptif

La plupart des théories adaptatives suppose que nous avons des références « internes », « intrinsèques » ou « mentales » (Delvaux and Soquet, 2007) à partir desquelles nos productions sont planifiées. Si la nature acoustique, motrice ou sensori-motrice de ces représentations est encore largement débattue, de nombreuses expériences ont démontré que le retour acoustique joue un rôle important dans le contrôle en ligne de la production de parole (Schwartz *et al.* ,2010). Avant de résumer les résultats des expériences mettant en valeur l'influence de la perception des productions d'autrui sur nos propres productions, nous allons passer en revue celles des expériences de perturbation du retour acoustique de nos propres productions.

Bauer *et al*.(2006) ont manipulé le volume du retour auditif de 20 sujets (10 hommes, 10 femmes) prononçant plusieurs fois le son [u] soutenu pendant 5s. Le gain du retour était soit augmenté soit diminué pendant 200ms de manière aléatoire pendant la production. La compensation observée est assez rapide (150ms) mais incomplète – de l'ordre de 1dB SPL pour des perturbations de 6dB. Cette compensation partielle a été observée dans d'autres études sur l'amplitude (Heinks-Maldonado and Houde, 2005).

Jones et Munhall (Jones and Munhall, 2000; Jones and Munhall, 2002) ainsi que d'autres auteurs (Burnett *et al.* ,1998; Bauer and Larson, 2003) ont montré qu'en présence de perturbation positive ou négative de f0, les sujets compensaient totalement cette perturbation, même si cette dernière était supposée non détectée (Jones and Munhall, 2002), inférieure au seuil de perception différentielle – Just Noticeable Difference (JND) – et que cette compensation était accompagnée d'un effet persistant après la disparition de la perturbation (« after effect »). Les temps de latence observés sont comparables à ceux cités ci-dessus : 210 ms en moyenne dans l'expérience I de Jones et Munhall (2002).

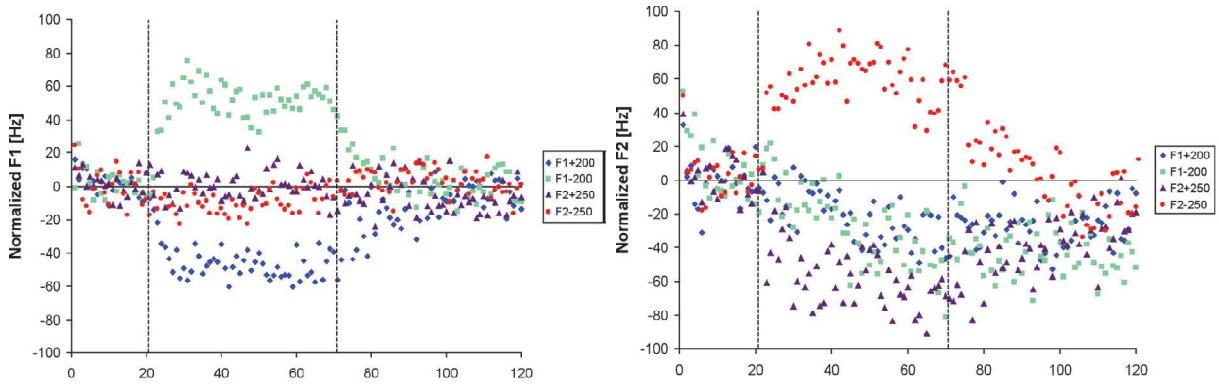


Figure 1. 3. Compensation des productions de voyelles dont les formants sont augmentés/diminués de plusieurs centaines de Hz avant d'être retournés au locuteur. A gauche : F1 ±100Hz ; à droite, F2±150Hz. On constate des compensations partielles – de l'ordre de 50% – des perturbations (d'après Purcell and Munhall, 2006).

Des travaux plus récents ont par ailleurs montré un phénomène analogue pour des perturbations du timbre de voyelles (Purcell and Munhall, 2006; Munhall *et al.* ,2009; MacDonald *et al.* ,2010; Cai *et al.* ,2011), ceci même si on demandait aux sujets de ne pas compenser cette perturbation (voir Figure 1. 3). Dans ce cas, les compensations sont partielles, de l'ordre de 50% de la perturbation.

1.4.1 Le How vs. What de Lindblom

Les perturbations plus écologiques de la communication parlée sont multiples : elles peuvent évidemment impliquer un environnement sonore réverbérant ou bruité par des sources sonores multiples, comprenant notamment d'autres sources vocales. Des indices indirects sur la « qualité » de nos productions sonores proviennent aussi de nos interlocuteurs, qui par des indices non verbaux, leurs demandes de confirmation ou les répétitions incorrectes ou approximatives de nos productions nous signalent que les signaux émis ne rencontrent pas leurs cibles linguistiques. Le retour des interlocuteurs est crucial pour la gestion de l'adaptation : Garnier *et al.* (2010) ont ainsi utilisé une tâche que les sujets ont accomplie seuls puis en collaboration avec un sujet de référence : ils

présentaient aux sujets une carte comportant 17 images de rivière avec des noms inventés de manière à récupérer un matériel phonétique adéquat. Les sujets devaient alors connecter les rivières entre elles par des flèches en suivant des règles précises. Ils ont remarqué que les sujets adaptaient leurs productions en condition bruitée (cocktail party) mais que cette adaptation était nettement plus marquée en situation d'interaction. Garnier *et al.* concluent donc que l'adaptation a bien un but communicatif.

A la suite de leurs travaux sur la variabilité, Lindblom *et al.* (Lindblom and Lindgren, 1985; Lindblom, 1987; Lindblom, 1990; Moon and Lindblom, 1994) ont d'abord proposé la théorie Hypo-Hyper (H&H) qui suppose que les locuteurs programment leurs productions en optimisant deux contraintes antagonistes : l'une, orientée locuteur, vise une économie gestuelle et favorise une coarticulation maximale où la distinctivité des sons est minimale ; l'autre, orientée interlocuteur, vise à augmenter les contrastes acoustiques de manière à faciliter le décodage du message par l'interlocuteur. Cette négociation est supposée être modulée par la prédictibilité du message linguistique : selon la théorie les mots fréquents ou facilement prédictibles par le contexte linguistique ou situationnel seraient hypoarticulés, alors que les mots difficiles, peu prédictibles ou supposés tels par le locuteur pour son interlocuteur, seraient hyperarticulés. Certains travaux menés sur l'incidence de la fréquence lexicale sur l'amplitude de la convergence phonétique (Goldinger, 1998) confirment le fait que cette dernière est moindre pour les mots les plus fréquents.

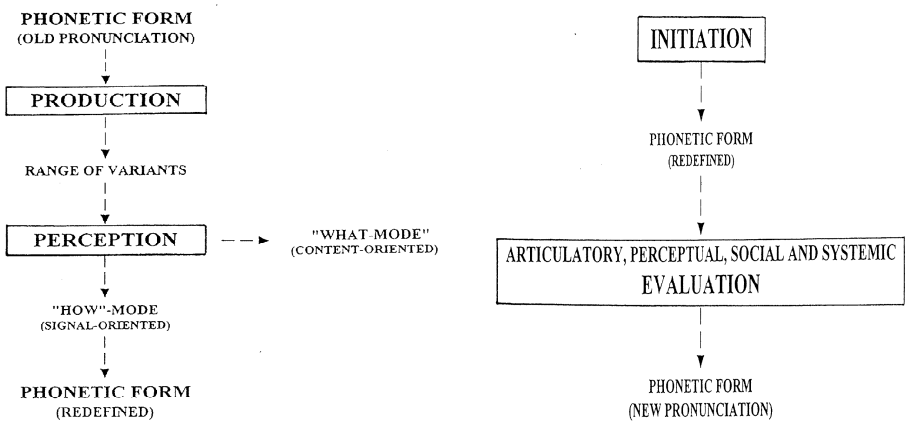


Figure 1. 4. Les deux étapes du changement phonétique selon Lindblom *et al.* (1995). A gauche, certaines productions réussissent à passer le filtre phonologique de l'auditeur et constituent une base potentielle de formes phonétiques candidates à l'élaboration de nouvelles prononciations. A droite, les prononciations peuvent se « fossiliser » et produire de nouvelles formes phonétiques incorporées au lexique du locuteur voire de son groupe social.

Dans leur modèle du changement phonétique, Lindblom *et al.* (1995) introduisent un modèle de la perception de parole à deux canaux ou deux modes (cf. Figure 1. 4), permettant à l'interlocuteur de porter son attention soit sur le message lui-même (la voie « *What »*) soit sur la manière dont le message est produit verbalement (la voie « *How* »). Selon les auteurs, la voie « *What »* est la voie privilégiée : elle sollicite le traitement des réalisations vocales par le filtre phonologique qui va « oublier » en quelque sorte la manière dont les traits phonologiques ont été implémentées. Elle traite donc des productions suffisamment canoniques, contrastées, régulières par rapport à la variabilité déjà apprise ou dont le contenu est suffisamment prédictible pour ne pas solliciter une attention particulière sur la structure phonétique du signal. En contraste et de manière plus marginale, certaines

productions – nouvelles, inattendues, difficiles ou peu prédictibles – vont passer le filtre phonologique et leurs formes phonétiques vont pouvoir être traitées et stockées de manière fine. Les « épisodes » phonétiques irréguliers sont, pour Lindblom *et al.*, à la base de la création de nouvelles formes phonétiques.

1.4.2 Les modèles de perception basés sur les exemplaires

A la fin des années 90, les modèles de perception basés sur les exemplaires (Goldinger, 1996; Johnson, 1997; Goldinger, 1998; Pierrehumbert, 2001) ont été proposés pour réconcilier perception catégorielle et sensibilité aux propriétés indexicales et autres attributs extralinguistiques. Les théories exemplaristes font l'hypothèse que les composantes phonologiques et extralinguistiques ne sont pas séparées dans les processus de perception, de représentation et de production de la parole, mais qu'elles sont accessibles simultanément du fait de la nature conjointe des représentations phonétiques et sémantiques stockées en mémoire.

Selon ces modèles, la convergence phonétique ne serait qu'une réaction automatique due au lien entre perception – qui résulterait d'une comparaison dynamique entre stimulus sensoriel et l'ensemble des traces collectées lors d'expériences, d'épisodes où actions, sensations, contexte et sémantique sont assimilées simultanément – et production – qui résulterait de l'activation des traces correspondantes. Ils postulent donc que, pendant une interaction, la perception élicite des traces qui sont stockées dans la mémoire à court terme et influencent les futurs productions par un mécanisme de rappel d'épisodes pertinents. Ceci n'exclut pas la perméabilité cognitive de ce processus de convergence automatique : on peut concevoir que des a priori situationnels – linguistiques, sociaux, culturels, etc. – préactivent/présélectionnent des exemplaires, qui seront alors utilisés de manière préférentielle pour ainsi montrer de manière consciente ou inconsciente ces mêmes a priori.

Ce lien entre production et perception a déjà été défendu par plusieurs théories. La théorie motrice de la perception de parole défendue par Liberman (Liberman *et al.* ,1967; Liberman, 1982; Liberman and Mattingly, 1985; Liberman and Mattingly, 1989) suppose que la perception d'un son active des représentations motrices, ces gestes articulatoires étant les « vrais » objets du décodage phonétique. La Théorie Réaliste Directe proposée par Fowler (1983; 1986; 1991) va même plus loin et propose que ses gestes sont accessibles directement à l'analyse sans médiation cognitive, inférences ou calculs complexes, l'acoustique jouant le rôle de la lumière dans la perception des objets physiques. Cette théorie de la perception directe, initiée par Gibson (1966), est en fait tout à fait compatible avec les théories exemplaristes, si on suppose que l'accès aux représentations sensori-motrices, indexicales voire sémantiques associée à chaque exemplaire sonore s'effectue indifféremment par l'une ou l'autre voire l'ensemble des représentations. A l'image d'un jeu de vecteurs multiparamétriques dont l'accès peut s'effectuer par une métrique spécifique à chaque composante : en effet, selon Goldinger (1996; 1998), la comparaison d'un stimulus à l'ensemble des traces – ou exemplaires, ce qui suppose un traitement, un classement, une sélection plus avancées des expériences sensori-motrices – stockées en mémoire génère un ensemble « d'échos » (1996; p.46) qu'Aubanel (2011) traduit comme « l'agrégat de toutes les traces activées, envoyé à la conscience depuis la mémoire à long terme ». Les principales critiques adressées à la perception directe (voir notamment Ohala 1986) sont de privilégier le « conduit vocal en mouvement » comme la cible de la perception et de renier toute influence cognitive dans cette médiation alors que les théories exemplaristes ne privilégient aucun accès et autorisent voire

stimulent un traitement cognitif permettant de pré-activer certaines traces ou représentations en fonction d'attentes ou d'amorçages contextuels.

Johnson (1997) propose que cette comparaison d'un stimulus acoustique avec l'ensemble des traces ou exemplaires acoustiques stockés en mémoire s'effectue de manière dynamique et temps-réel à la manière d'une programmation dynamique ou d'un alignement par modèles statistiques plus complexes (voir par exemple le décodage en ligne de modèles HMM proposé et évalué par Bloit J. (2008). Johnson postule de plus que cet accès s'effectue sans normalisation préalable. Les exemplaires sont stockés tels quels et portent donc toutes les caractéristiques paralinguistiques. L'interlocuteur a donc un accès immédiat à des caractéristiques idiosyncratiques liées à la voix, à l'accent, aux circonstances d'écoutes précédentes. Ces traces peuvent donc ainsi déclencher des échos permettant d'accéder à des informations plus élaborées sur son interlocuteur comme ses variables dialectales voire non-linguistiques comme son âge, son genre ou son état émotionnel.

On peut alors se demander comment ces traces puis ces modèles linguistiques, phonologiques et phonétiques des interlocuteurs peuvent influencer la production des cibles linguistiques du locuteur.

1.4.3 Les liens perception/action

Mis en évidence en 1992 par Rizolatti et ses collègues de l'Université de Parme en Italie (di Pellegrino *et al.* ,1992; Rizzolatti and Arbib, 1998; Rizzolatti and Craighero, 2004), les neurones miroirs sont des neurones aux caractéristiques assez singulières situés dans l'aire F5 du cortex prémoteur ventral du singe. Ces neurones s'activent non seulement lorsque l'animal effectue un mouvement intentionnel mais aussi lorsque l'animal voit un de ses congénères ou un expérimentateur faire ce même mouvement particulier avec les mêmes moyens et dans le même but : aucun déclenchement des neurones activés en action et en observation de l'action si l'autre utilise un outil pour la même action ou si le geste a une autre finalité. La réponse de ces neurones-miroirs est donc fortement contingente et liée à l'expression de l'intentionnalité du geste observé. La compréhension des actions effectuées par autrui reposerait ainsi sur la « résonance » de ces actions dans le système moteur de l'observateur via son expérience sensori-motrice.

Chez des singes de laboratoire, Kohler *et al.* (2002) ont ainsi trouvé des neurones-miroirs audiovisuels qui déclenchaient aussi bien à l'observation d'actions sonores (ouvrir une cacahuète, déchirer un morceau de papier, etc.) qu'à l'écoute des sons associés à ces actions. De même, Ferrari *et al.* (2003) ont trouvé des neurones-miroirs « neurones-miroirs transitifs de la bouche », décharge lorsque le singe exécute et observe des actions buccales d'ingestion d'objets (e.g. saisir, mordre, lécher) faites par l'expérimentateur, le singe ou ses congénères, alors que d'autres neurones répondent à l'exécution et à l'observation d'actions de succion. Les neurones-miroirs semblent donc dépendre de l'organe sollicité par l'action : on a effectivement identifié des neurones-miroirs du pied ou de la main (Binkofski et Buccino, 2004 ; Binkovski *et al.*, 1999 ; Ehrsson *et al.*, 2000 ; Buccino *et al.*, 2001).

C'est ce qui a amené ces mêmes chercheurs (Rizzolatti and Arbib, 1999; Rizzolatti and Sinigaglia, 2006) à penser que les neurones miroirs pourraient nous éclairer sur les fondements cognitifs du langage en constituant le substrat neuronal de notre capacité à comprendre la signification des actions d'autrui qui fonde toutes les relations sociales (voir notamment les théories de l'esprit proposées par

Baron-Cohen *et al.* ,1985; Baron-Cohen, 1994; Leslie, 1994; Baron-Cohen *et al.* ,1997). Ce système de correspondance entre perceptions, actions et intentions nous aiderait à attribuer des états mentaux à autrui et à interpréter leurs actions comme des comportements intentionnels. Il est ensuite facile d'imaginer comment ce mécanisme d'interprétation de la gestualité ait initié une communication gestuelle puis une communication verbale qui aurait exaptée le conduit vocal de sa fonction première de respiration, mastication et déglutition. Cette primauté du geste au verbe, avancée par Tomasello *et al* (Call and Tomasello, 2008; Tomasello, 2008), s'oppose à l'argumentaire développé par Rizzolati et Arbib (1998) dans lequel la communication brachio-manuelle vient au contraire compléter les gestes orofaciaux lors d'étapes-clés de la maturation d'un système de communication verbal, notamment pour décrire, qualifier et désigner (Tomasello and Kruger, 1992), donc lexicaliser les productions sonores.

Quelle que soit la théorie de l'évolution que la postérité retiendra, les neurones-miroirs offrent un substrat neuronal pour une faculté primordiale de l'homme, plus limitée chez les primates: l'imitation. L'imitation (voir §1.1) est la capacité d'un individu à exécuter une action d'abord en l'observant puis à apprendre à exécuter cette même action par lui même. Cette capacité exerce un rôle central dans le comportement humain, tant dans l'apprentissage moteur que dans la mise en place de la communication et des compétences sociales (Piaget, 1962; Tomasello *et al.* ,1993). Ce stockage de l'action, de son but, de son sens et de ses conséquences sensorielles permettant cette mise en résonnance de l'observation d'actions d'autrui avec ses propres représentations, serait donc « entaché » de caractéristiques de l'action, du but, du sens et des conséquences sensorielles attribuées par notre congénère à son comportement.

1.4.4 Parole et imitation

1.4.4.1 Imitation involontaire

Dans le cadre de l'étude des liens entre gestes et parole, Gentilucci *et al.* (Gentilucci *et al.* ,2001; Gentilucci, 2003; Gentilucci *et al.* ,2004) ont conduit un ensemble d'expériences montrant que l'articulation de sons isolés, de syllabes voire de mots est influencée par la manipulation simultanée d'objets par le locuteur ou par son interlocuteur : la puissance vocale maximale des productions vocales – ou plutôt l'ouverture de la mâchoire – est proportionnelle à la taille de l'objet manipulé ou que l'on voit manipuler. Ce lien fonctionnel entre langage et motricité semble réciproque : saisir avec la main semble activer un programme moteur de saisie avec la bouche – parler – et réciproquement. Il est donc logique que l'observation du geste vocal d'autrui influence a fortiori nos gestes vocaux. Gentilucci et Bernardis (2007) ont ainsi mis en place trois tâches de répétitions de VCV ((/aba/, /ada/ et /aga/) avec les modalités audio, visuelle, puis audio-visuelle. Ils ont mesuré plusieurs paramètres (ouverture/fermeture des lèvres, f0, F1, F2, intensité, durée) pendant un enregistrement de référence puis les trois tâches et a remarqué que, lorsque la photo du sujet de référence (un homme) était présente, les sujets (des femmes) adaptaient leurs ouverture/fermeture des lèvres à celles du sujet de référence. Il a également observé une diminution significative des paramètres acoustiques. Les sujets utilisaient donc tous les signaux disponibles pour se rapprocher de leur interlocuteur. On peut noter que, dans une deuxième expérience, Gentilucci et Bernardis ont voulu voir si le genre du sujet de référence allait également être un facteur pouvant influencer l'adaptation.

Il est intéressant de comparer l'imitation involontaire avec l'imitation volontaire pour laquelle les effets observés sont souvent comparables à ceux de l'imitation involontaire mais avec une amplitude plus forte.

1.4.4.2 Imitation volontaire

Sato *et al.* (Sato *et al.* ,2010; Sato *et al.* ,2011) ont comparé une tâche de répétition de voyelles avec une tâche d'imitation de voyelles. Ils voulaient démontrer que les sujets allaient imiter de manière automatique les stimuli acoustiques pendant la tâche de répétition et que par conséquent l'effet serait similaire à celui trouvé pendant la tâche d'imitation mais avec une amplitude moins élevée. Pour caractériser cet effet, ils ont étudié les variations du f0 et du F1 entre la prononciation des voyelles de référence (en condition lue) et celle pendant les tâches de répétitions et d'imitation. Ils ont effectivement observé une adaptation significative du f0 et du F1 qui était plus forte pendant la tâche d'imitation. Ils ont également voulu tester si l'adaptation restait après la tâche et ont donc enregistré de nouveau les sujets en condition lue. Ils ont bien remarqué une persistance de l'adaptation que l'on peut aussi nommer « After-effect ». Ce phénomène « d'After-effect » confirme que des traces ont été stockées pendant la tâche et que celles-ci ont conditionné les productions suivantes.

La faculté d'imitation volontaire de voyelles – on parle aussi de « shadowing » – a été introduite par Chistovitch *et al.* (1966a; 1966b). Ils ont demandé à des sujets d'imiter des voyelles synthétiques créées à partir d'un continuum entre les voyelles [i]-[ɛ]-[a]. Ils ont remarqué une imitation catégorielle des sujets qu'ils ont attribuée à une représentation discrète des voyelles utilisée d'abord pour guider l'articulation puis pour faciliter le stockage des voyelles dans la mémoire à long-terme. Kent (1973) a répliqué la même expérience mais en étudiant deux continuum différents, un de [u] à [i] et l'autre de [i] à [æ]. Il a également observé une imitation catégorielle mais distincte entre les deux continuums. Plus de catégories pavant le continuum entre [i] à [æ], la transition de [i] à [æ] présentait plus d'attracteurs que celle de [u] à [i]. Repp et Williams (1985; 1987) ont conclu de leurs études que l'imitation catégorielle pouvait provenir de différents facteurs comme les contraintes articulatoires, les habitudes de prononciation des sujets mais que cette imitation catégorielle s'organisait autour des espaces propres à chaque son (voir la théorie quantique de la parole proposée par Stevens 1972). Ils ont utilisé le même paradigme que leurs prédécesseurs mais en utilisant pendant la première expérience des voyelles synthétiques et dans la seconde des voyelles produites par leurs sujets – les sujets pouvaient donc en théorie imiter toutes les voyelles –. Ils ont remarqué que dans le second cas l'imitation était plus précise mais toujours catégorielle.

1.4.5 Commentaires

Gentilucci a étudié si le genre des sujets allait influencer le phénomène d'adaptation. Dans ce cas, on ne peut plus utiliser le terme « processus automatique » pour définir l'adaptation car des facteurs sociaux (ici le genre) modulent l'amplitude de la convergence. Ils ont cependant juste observé un effet sur l'intensité mais aucun effet sur les autres paramètres.

1.5 L'adaptation communicative

Giles *et al*. (Giles and Clair, 1979; Giles *et al.* ,1987; Giles *et al.* ,1991) ont développé tout d'abord une théorie de l'accommodation en parole (« Speech Accommodation Theory » ou SAT) (Giles, 1973)

qui postule que, pendant une conversation, les interlocuteurs modifient leurs caractéristiques vocales pour atteindre divers buts communicatifs (Giles, 1973; Giles and Clair, 1979); puis une théorie appelée Théorie de l'Accommodation Communicative (« Communication Accommodation Theory » ou CAT) (Giles *et al.* ,1987) qui propose que, pendant une interaction, les locuteurs utilisent le langage comme un outil soit pour diminuer la distance sociale qui les sépare, on parle alors de convergence :

« *it is probably safe to assume that these shifts resulted in a favorable appraisal of the speaker, that is, they have created an impression that the speaker is trying to accommodate to his or her listener(s)* » (Giles and Clair 1979, p.47)

soit pour accentuer cette différence, on parle alors de divergence **:**

« *speech divergence may be an important strategy for making oneself psychologically and favorably distinct from outgroup members* » (Giles and Clair 1979, p.52)

Les buts de l'adaptation peuvent être multiples :

- simplifier l'échange de messages dont le contenu est très dépendant du contexte (Lakin *et al.* ,2003)
- améliorer la capacité de percevoir, comprendre, accepter de nouvelles informations. (Traum and Allen, 1992; Allwood, 2002)
- aider à atteindre des buts communs (Clark, 1996) et améliorer la qualité de l'interaction (Babel , 2009)
- contribuer à la compréhension mutuelle en diminuant la distance sociale (Babel , 2009)
- signaler à l'interlocuteur que l'on souhaite poursuivre l'interaction (Labov, 2001)
- permettre de faire émerger des formes plus stables au sein d'une communauté (Garrod and Doherty, 1994). Il faut cependant noter que ces formes sont diverses : ainsi différents gestes de pointage émergent suivant les cultures (voir l'utilisation de l'index, de la main, des lèvres et du regard dans Wilkins, 2003).

Cette théorie est cohérente avec l'idée de Tajfel & Turner (1979) qui affirme que la convergence est un outil social qui permet aux individus de déterminer leur appartenance à un groupe en se comparant aux autres groupes existants. Dans ce cas, un individu évaluerait positivement le groupe auquel il appartient. La CAT explique comment les personnes vont alors modifier leurs productions verbales pour essayer de maintenir cette distinction entre les groupes. Cette capacité de rapprocher (ou éloigner) leurs productions de celles de leur interlocuteur permettent aux individus de se placer dans un espace conversationnel et plus largement dans un espace social. Les locuteurs peuvent alors adapter leurs productions en fonction de deux niveaux :

- Les facteurs sociaux qui caractérisent les individus ou les groupes auxquels ils appartiennent
- Le contexte social dans lequel l'interaction se situe

On peut alors considérer que la convergence phonétique fait partie d'un ensemble de stratégies utilisées par les individus pour atteindre des buts particuliers tels que l'approbation ou encore

l'intégration. La convergence phonétique est donc étroitement liée avec les variables sociales tels que le statut ou la dominance.

La CAT ne prédit pas seulement un rapprochement des locuteurs en interaction. Elle propose également trois autres types de comportement : le maintien, la complémentarité et la divergence. Les individus vont avoir tendance à diverger lorsqu'ils vont vouloir accentuer la distance sociale qui les sépare comme face à une menace. S'ils ne modifient pas leurs productions, on parle de maintien et s'ils utilisent un caractère propre non contrôlé de leur langage pour accentuer la distance sociale on parle alors de complémentarité, par exemple, les hommes vont avoir tendance à parler de manière encore plus grave avec des femmes et pas avec des hommes. Ces types de comportements ne sont pas exclusifs les uns des autres et peuvent être dynamiquement utilisés pendant une interaction (de Looze *et al.* ,2011).

Dans les parties suivantes, nous allons présenter les études qui ont portées sur la convergence inter-dialectale puis celles sur la convergence en interaction.

1.5.1 Contacts linguistiques

Des études portant sur des personnes bilingues ont démontré que la langue natale avait une influence sur les productions en langue seconde (Flege, 1987; Flege and Eefting, 1987; Bullock *et al. ,* 2006) mais que cette influence diminuait au cours de l'apprentissage, les sujets apprenant à faire la distinction entre les deux systèmes phonologiques ou assimilant les nouveaux phonèmes comme des variantes de ceux de leur langue d'origine ("equivalence classification" selon Flege and Eefting, 1987). Ainsi, au début de leur apprentissage, les sujets ne possédant pas de traces phonétiques de la nouvelle langue, ils utilisent inconsciemment le modèle de leur langue natale.

Fowler *et al.* (Sancier and Fowler, 1997; Fowler *et al.* ,2008) ont montré que l'environnement (i.e. le pays de résidence) avait une influence sur les productions des sujets bilingues. Ils ont analysé les productions d'un sujet de référence bilingue en Anglais et en Portugais et ont remarqué que s'il rentrait d'un séjour au Brésil alors ses productions en Anglais étaient influencées par sa pratique du Portugais et réciproquement. Chang (2011) a remarqué que des apprenants anglais qui débutaient en Coréen avaient tendance à diminuer leur premier formant pour se rapprocher de celui du Coréen en produisant des phonèmes en Anglais. L'environnement social dans lequel se trouvaient les sujets (stage intensif de Coréen, vie en communauté avec des personnes coréennes) influençaient donc leurs productions. De la même manière, Cibelli (2009) a observé un pattern d'adaptation entre l'Espagnol et l'Anglais qui était asymétrique. Les sujets adaptaient la durée et le F3 des voyelles étudiées à celui du sujet de référence qui parlait Anglais avec un accent espagnol mais pas avec l'autre qui parlait Espagnol avec un accent Anglais. Cette asymétrie peut s'expliquer par le fait que les sujets exclusivement féminins s'adaptaient au sujet de référence qui était du même sexe qu'elles. Kim *et al.* (2011) ont voulu observer l'amplitude de la convergence en fonction de la distance linguistique qui séparait les sujets. Pour cela, ils ont fait participer à une tâche de collaboration des paires de même dialecte, de même langue et de langues natales différentes. Ils ont alors observé une plus grande convergence (environ 60% de similarité) pour les sujets de même dialecte donc ceux pour qui la distance linguistique était la plus courte. La charge cognitive pour accomplir la tâche dans une langue qui n'était

pas la langue d'origine des sujets devaient être trop élevée ainsi ils ne se concentraient pas sur l'aspect communicatif de l'interaction et n'utilisaient donc pas les diverses stratégies d'adaptation.

Nous allons maintenant présenter les études qui ont été faites entre plusieurs dialectes. Pour ces études, on a donc diminué la distance linguistique pour étudier la convergence entre les sujets.

1.5.2 La convergence inter-dialectale

Delvaux et Soquet (2007) ont étudié le phénomène de convergence au sein d'une même langue mais en utilisant des dialectes différents, on peut alors parler de convergence phonologique. Ils ont comparé les productions de segments critiques (/ɔ/, /i/ et /ɛ/) en Flamand et en Wallon en utilisant une tâche de description en perturbation ambiante. Ainsi pendant la tâche les sujets entendaient les productions des personnes de l'autre régiolecte et Delvaux et Soquet ont analysé les effets obtenus. Ils ont utilisé les coefficients MFCC (Mel-Frequency Cepstral Coefficient), la durée des segments critiques et leurs trois premiers formants pour caractériser la convergence. Contrairement à Kim *et al*. (2011), ils ont observé une adaptation des productions des sujets à celles de l'autre régiolecte mais la charge cognitive était moins élevée puisque les sujets participaient juste à une tâche de description. L'effet observé reste cependant faible puisque suivant les segments critiques étudiés, ils ont obtenu entre 10 et 25 % de convergence (voir Figure 1. 5).

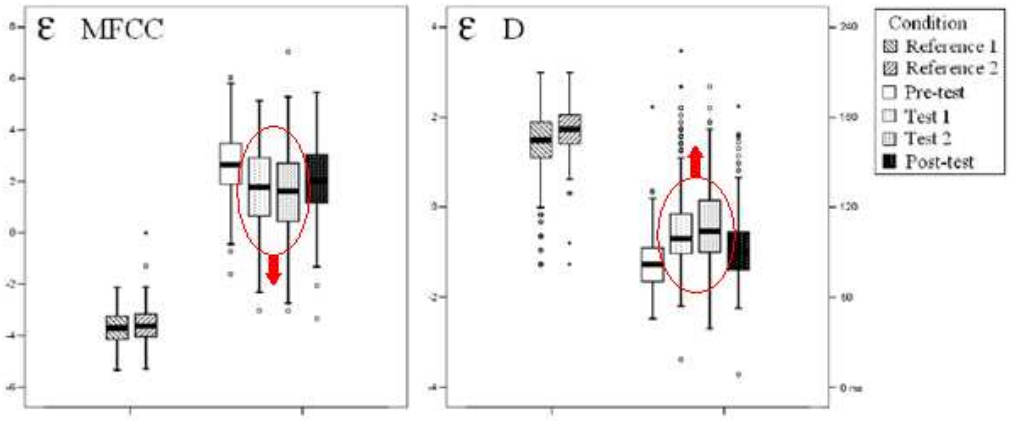


Figure 1. 5. Boxplots représentants les scores des fonctions discriminantes pour les MFCCs et la durée du /ɛ/ pour chaque régiolectes et chaque condition

Babel (2008; 2009) a introduit le facteur social dans une tâche de répétition de voyelles ([i], [æ], [ɑ], [o] et [u]) en incluant la photo des deux sujets de référence qui sont de type caucasien et de type africain. Elle a également testé des sujets masculins et féminins pour observer si les hommes et les femmes adoptaient des comportements d'adaptation différents. Elle a demandé à ses sujets de noter l'attractivité des sujets de référence sur une échelle de 1 à 10 et a également utilisé la tâche d'association implicite (IAT) définie par Greenwald *et al.* (1998) pour quantifier l'attractivité des sujets de référence. Elle a caractérisé la convergence en utilisant les premier et second formants et a comparé les patterns de convergence trouvés en fonction de l'attractivité du sujet de référence et du sexe des sujets. Elle a remarqué que les hommes et les femmes n'utilisaient pas les mêmes stratégies d'adaptation. Les femmes avaient tendance à se rapprocher des sujets qu'elles trouvaient attractifs

alors que les hommes s'en éloignaient. De plus, elle a uniquement observé une adaptation pour les voyelles [æ] et [ɑ].

Nous allons maintenant passer en revue les études qui traitent de la convergence en interaction pour des sujets qui partagent le même régiolecte.

1.5.3 La convergence en interaction

D'autres chercheurs ont mis en place des paradigmes utilisant des tâches de collaborations forçant ainsi la coopération et l'attention mutuelle des sujets. Ils ont ainsi poussé les sujets à utiliser des stratégies de communication dont la convergence pour réussir la tâche qui leur était présentée. Pardo (2006) a observé l'adaptation entre des paires de même sexe en utilisant un concept développé au Human Communication Research Center de l'Université de Glasgow et Edinbourg nommée « Map Task » (Anderson *et al.* ,1991). Les sujets devaient collaborer pour reproduire un chemin sur une carte et prononçaient ainsi plusieurs fois les mots cibles représentés par des icônes. Pardo a ajouté un facteur supplémentaire à son étude en ne donnant pas aux sujets le même rôle pendant l'interaction. Un des sujets connaissait la carte à reproduire et devait la décrire au second pour que celui-ci la reconstitue. Pour caractériser la convergence elle a utilisé un test de perception et a alors souligné un phénomène de convergence très dépendant du sexe et du rôle des sujets. Les résultats obtenus montraient un taux de similarité qui oscillait entre 55 et 75% (Figure 1. 6). Cependant le test de perception mis en place par Pardo a été créé à partir de stimuli enregistrés dans des conditions différentes (conditions lues vs. parlées), ainsi les juges ont pu détecter une convergence des paramètres d'enregistrements plutôt qu'une adaptation des sujets à leur interlocuteur.

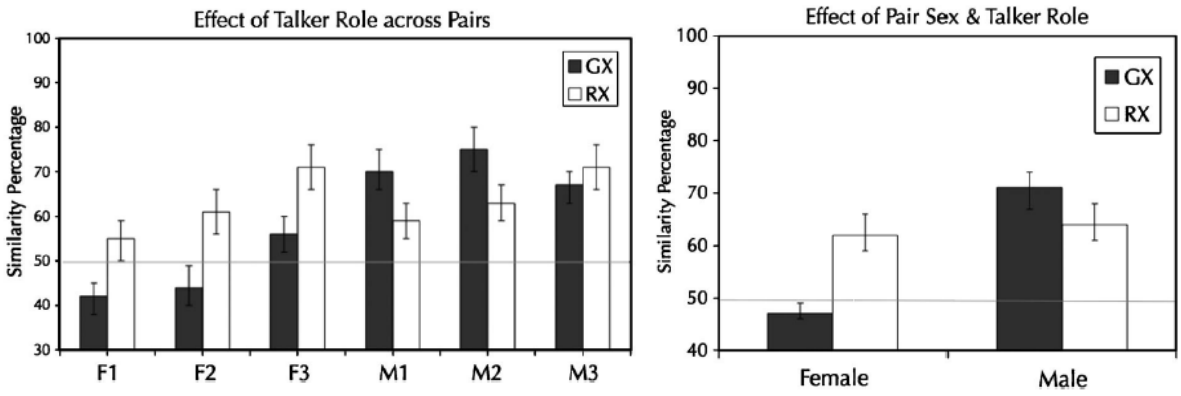


Figure 1. 6. Résultats du test AXB mené par Pardo. La barre noire correspond au taux de convergence du donneur vers le receveur et la barre blanche correspond au taux de convergence du receveur vers le donneur

Aubanel (Aubanel and Nguyen, 2010; Aubanel, 2011) a, comme Delvaux et Soquet, étudié l'influence de deux régiolectes, le Français standard et le Français méridional mais en interaction. Il a demandé à des paires de même sexe et de même rang social – basé sur l'échelle de Désirabilité Sociale de Crowne et Marlowe (1960) – d'associer des photographies à des noms de famille inventés de manière à récupérer le matériel linguistique approprié à l'analyse de la convergence puis de créer des groupes avec ces photographies. Les analyses ont été faites sur des segments critiques typiques des deux régiolectes en utilisant deux méthodes, un classifieur de Bayes et une analyse linéaire discriminante sur les enveloppes spectrales réduites par DCT (Discrete Cosine Transform). Ils ont

remarqué que le phénomène de convergence était fortement dépendant des paires, des segments critiques et des jeux étudiés (voir Figure 1. 7). De plus les effets observés restaient très faibles : la dynamique idéale de convergence illustrée par la Figure 1. 7 n'est obtenue que pour deux des dyades.

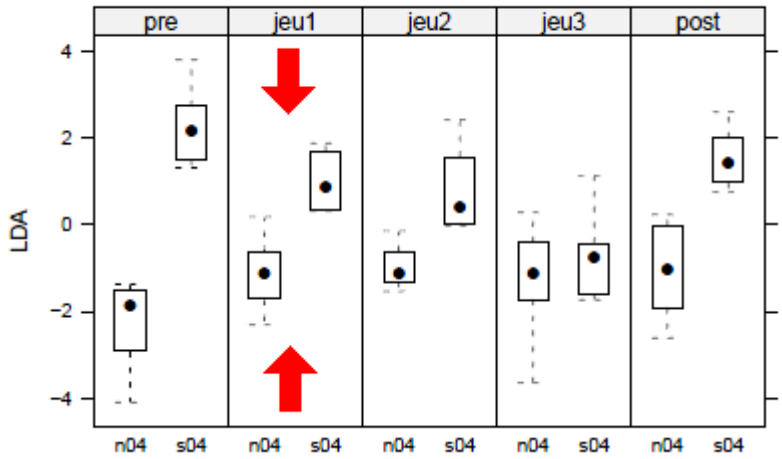


Figure 1. 7. Score discriminant associé aux réalisations d'un segment des deux locuteurs d'une dyade au cours des 5 phases des enregistrements : pré-test, jeu 1, jeu 2, jeu 3, post-test.

La manière de caractériser la convergence sur des segments critiques pour distinguer des différences phonologiques est moins exigeante que pour la convergence phonétique. En effet, il suffira de collecter le nombre de segments critiques attribués à chaque variante phonologique en pré-test et de comparer ce nombre avec celui obtenu de la même manière pendant l'interaction. Pour l'analyse de la convergence phonétique, il est nécessaire de trouver une manière fiable et objective (i.e. une métrique) capable de comparer les prononciations des phonèmes entre les sujets.

Kousidis *et al*. (Kousidis *et al.* ,2008) et Gregory *et al*. (Gregory and Hoyt, 1982; Gregory 1986; Gregory and Webster, 1996) ont d'étudié le phénomène de convergence dans une tâche conversationnelle non scénarisée. Cette condition est la plus écologique que l'on puisse trouver mais elle pose également beaucoup plus de problèmes pour l'analyse de données car elle ne garantit pas une représentativité suffisante des phonèmes critiques. Kousidis a remarqué seulement un effet de convergence rapide principalement sur l'intensité et le débit de parole non cumulatif avec le temps. Aucun effet significatif n'a été observé sur le pitch. L'étude de ce paramètre est d'autant plus difficile en interaction au titre que les sujets utilisent différents types de phrases (déclarative, interrogative, exclamative). Cependant, Gregory *et al.* ont observé une adaptation du pitch pour ses sujets qui était très dépendant du rapport social entre les interlocuteurs. Ainsi son sujet de référence (Larry King) a interviewé 25 célébrités et a adopté des comportements différents selon que son invité soit de rang social plus élevé ou moins élevé que le sien. Il s'adaptait aux personnes de rang social plus élevé alors que les invités de rang social moins élevé s'adaptaient à lui.

1.6 Discussion

Tous ces exemples ont démontré que le phénomène de convergence phonétique ne pouvait pas s'expliquer complètement avec la CAT ou avec les modèles à exemplaires. En effet, d'un côté, la CAT ne peut pas expliquer ce qui va se passer au tout début d'une interaction entre deux inconnus car le

contexte social n'est pas encore défini. De l'autre côté, il est évident que les facteurs sociaux vont avoir une importance grandissante au cours de l'interaction ainsi le phénomène d'adaptation ne serait pas uniquement automatique.

Notre hypothèse est donc la suivante (cf. Figure 1. 8). Pour qu'il puisse y avoir une convergence phonétique entre deux personnes qui se rencontrent, il faut qu'elles collectent d'abord, toutes les deux, des traces des productions de leur interlocuteur et, à partir de ce moment, ces traces vont automatiquement ajuster les productions des personnes vers celles de leur interlocuteur. Le phénomène de convergence phonétique peut être expliqué par les modèles à exemplaires qui expliquent comment une adaptation peut être réalisée à partir de peu d'exemples : Ostry *et al.* (Malfait *et al.* ,2005) ont ainsi montré que leurs sujets apprenaient rapidement – quelques essais – et de manière durable à adapter leurs mouvements d'extension du bras perturbés de manière inopinée par des forces externes, mais que ces contrôles compensatoires étaient locaux et non généralisables (Mattar and Ostry, 2007).

Ainsi une fois qu'assez d'exemplaires ont été perçus, les locuteurs construisent un modèle de leur interlocuteur et peuvent alors faire appel à ce modèle pour adopter une stratégie d'adaptation cohérente avec le contexte social de l'interaction. Ici, c'est donc la CAT qui va décrire le phénomène de convergence phonétique. Callan (2004) a démontré que l'activité cérébrale était moins importante pour des voyelles très courantes par rapport à des voyelles plus rares. Cela peut se traduire par le fait que le cervelet crée un modèle interne des voyelles rares au fur et à mesure de leurs productions. Ainsi, lorsque les modèles internes ne sont pas encore créés, donc au début de l'interaction, la charge cognitive est trop importante donc les locuteurs s'adaptent de manière automatique. Quand les modèles internes ont été définis, il y a une réallocation de l'activité cérébrale (Sato *et al.* ,2011) et les sujets peuvent alors déterminer la stratégie d'adaptation qu'ils veulent utiliser face à leur interlocuteur.

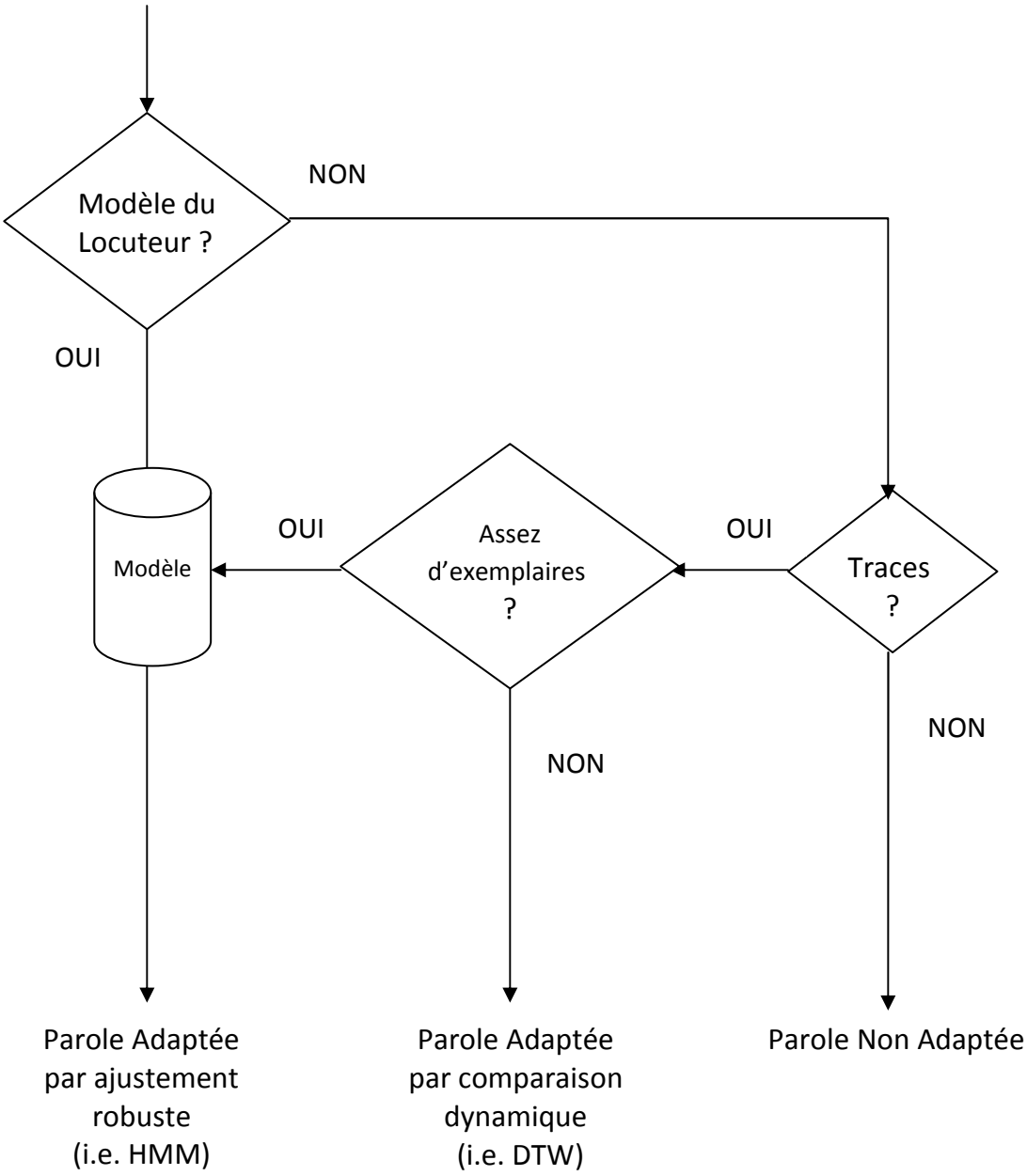


Figure 1. 8. Processus d'adaptation en interaction selon la quantité d'exposition préalable aux productions de l'interlocuteur. Le processus de comparaison et la réaction du locuteur à un stimulus dépend de la quantité d' « exemplaires » sonores auxquels il a déjà été confronté.

1.7 L'adaptation en interaction Homme-Machine

Les interactions hommes-machines étant de plus en plus utilisées, des chercheurs ont étudié comment les personnes réagissaient face à ces systèmes. Ils ont tenté de voir si les mêmes stratégies étaient utilisées pour des interactions hommes-hommes et hommes-machines.

D'après Branigan *et al.* (2000; 2010), l'étude de l'alignement en interaction homme-machine est importante pour deux raisons. Tout d'abord, cela permettra d'améliorer les systèmes de dialogue en les rendant plus naturels mais aussi d'étudier leurs limites. Dans un deuxième temps, cela aidera à comprendre davantage l'alignement en interaction homme-homme grâce à un contrôle de paramètres

plus facile pour les paradigmes. Différents types d'alignement ont été observés pendant ce type d'interaction. Oviatt *et al.* (1998; 2004) ont remarqué que leurs sujets hyperarticulaient et parlaient plus doucement pour être sûrs d'être « compris » par l'ordinateur alors que Branigan *et al.* ont essentiellement observé un alignement au niveau lexical.

L'alignement du système de dialogue en interaction homme-machine peut se faire à différents niveaux. Van Vugt *et al.* (2010) ont d'abord démontré que les personnes préféraient interagir avec un agent conversationnel animé qui leur ressemblait physiquement (même si ils n'avaient pas remarqué la ressemblance) et plus particulièrement si cet agent était perçu comme serviable. Bailenson et Yee (2005) ont trouvé que leurs sujets préféraient interagir avec des agents conversationnels animés qui imitaient leurs gestes de la tête. De leur côté, Fogg et Nass (1997) ont observé que les sujets travaillaient mieux avec un système qui les avait précédemment aidés pour tâche. Nass et Lee (2001) ont étudié l'effet de l'alignement de la personnalité du système de dialogue – introverti ou extraverti – avec celle des sujets en jouant sur les caractéristiques de la voix. Les auteurs ont remarqué que les sujets préféraient des voix qui correspondaient à leur personnalité. Enfin, Ward et Nakagawa (2002) ont remarqué que sujets jugeaient plus favorablement un système de téléphonie qui adoptait leur débit de parole.

Dans tous les cas présentés précédemment, l'alignement du système de dialogue inspire une impression positive au sujet ce qui va le pousser à s'adapter encore plus à son « partenaire ». Cette observation confirme le fait que les systèmes de dialogue sont considérés comme des agents sociaux par les sujets et que ces derniers vont donc adapter leur comportement comme pour les interactions hommes-hommes (Nass and Moon, 2000).

Il semble que les sujets vont systématiquement adapter leur manière de parler au système de dialogue. Zoltan-Ford (1991) a comparé deux modes de communication hommes-machines, une écrite et une orale, en faisant varier le vocabulaire et la longueur des énoncés. Elle a remarqué que les utilisateurs du système adaptaient la longueur de leurs phrases et leur vocabulaire en fonction de ceux du système en condition orale mais pas en condition écrite et que les sujets étaient plus ouverts aux systèmes hommes-machines après avoir interagir avec le système de manière orale. Des chercheurs ont également observé une adaptation d'autres paramètres acoustiques avec ceux de systèmes de dialogue comme l'intensité (Coulston *et al.* ,2002) ou le débit de parole (Bell *et al.* ,2003). Oviatt *et al* (2004) ont étudié si les enfants s'adapter à un système de dialogue en faisant varier les propriétés acoustiques de la voix de synthèse. Ils ont remarqué que les enfants s'alignaient avec leur partenaire virtuel (par exemple pour l'amplitude et les pauses). Ce même phénomène a été observé avec des adultes par Suzuki et Katagiri (2007).

Si on compare maintenant l'amplitude de l'alignement entre les interactions hommes-machines et les interactions hommes-hommes, on remarque que l'adaptation ne se fait pas pour les mêmes raisons et ne va pas évoluer de la même manière. Branigan *et al.* (2003) ont remarqué que les sujets avaient plus tendance à s'aligner avec un système de dialogue plutôt qu'avec un interlocuteur humain. Ils attribuaient cet effet au fait que pendant une interaction homme-machine le but principal de l'alignement est le succès de la communication et qu'il était fortement conditionné par la première impression du sujet sur le système de dialogue. Par exemple, Pearson *et al*. (2006) ont montré que des sujets s'alignaient plus avec un système de dialogue qui leur était présenté comme basique. En interaction homme-homme, l'alignement est davantage conditionné par l'affect social et évolue donc pendant l'interaction en fonction du contexte.

Si on s'intéresse maintenant à un aspect plus applicatif, Ward et Litman (2007) ont étudié le lien entre la convergence lexicale et la convergence de paramètres acoustiques avec la facilité d'apprentissage face à un *Système Tuteur* Intelligent (ITS en Anglais). Ils ont remarqué que le phénomène de convergence était un bon indicateur de l'apprentissage : plus les étudiants convergeaient avec le tuteur, meilleur était leur apprentissage. Une convergence des sujets pouvait traduire une implication des étudiants dans l'interaction avec le tuteur et ainsi une volonté d'apprendre ce qu'il leur enseignait.

Toutes ces études démontrent que le phénomène de convergence est un aspect essentiel de la communication et particulièrement important pendant les interactions face-à-face qu'elles soient entre humains ou avec un agent conversationnel animé. Il paraît donc crucial de bien comprendre toutes les implications de ce phénomène. Nous allons donc pour cela définir un paradigme qui nous permettra d'étudier le phénomène de convergence phonétique en interaction face-à-face.

Articles	Théorie	Méthode	Caractérisation	Résultats
(Sancier and Fowler 1997)	CAT	Production en Anglais/Portugais pour des personnes bilingues	VOT de [p], [t] et [k]	VOT dépendant du dernier séjour du sujet de référence
(Fowler *et al.* 2008)	CAT	Production en Anglais/Français pour des personnes bilingues	VOT de [p], [t] et [k]	VOT plus long en Français pour les sujets bilingues
(Chang 2011)	CAT	Apprentissage du Coréen par des Anglophones	F1, F2	Influence de la L2 sur la L1 au début de l'apprentissage
(Cibelli 2009)	CAT	Production en perturbation ambiante en Anglais/Espagnol pour des personnes bilingues	F1, F2, F3, durée	Adaptation de la durée et du F3 avec un sujet de référence parlant Anglais avec un accent Espagnol
(Kim *et al.* 2011)	CAT	Espace partagé + Production en Anglais/Coréen pour des personnes bilingues	Test XAB pour comparer début et fin d'interaction	60% de similarité avec le signal de fin d'interaction
(Delvaux and Soquet 2007)	CAT	Production de mots cibles + Perturbation ambiante	Analyse linéaire discriminante sur 20 premiers coefficients MFCC, F1, F2, F3	Environ 20% de convergence sur les MFCCs
(Babel 2009)	CAT	Tâche de répétition de voyelles	Distance euclidienne pour F1 et F2	Convergence pour [a] et [æ]. Patterns différents selon le sujet de référence et le sexe des participants.
(Pardo 2006; Pardo *et al.* 2010)	CAT	Production de mots cibles + Espace partagé	Test AXB	60% de similarité avec le donneur pour les femmes et 70% de similarité avec le receveur pour les hommes
(Aubanel and Nguyen 2010; Aubanel 2011)	CAT	Production de mots cibles + Espace partagé	Classifieur de Bayes, Analyse linéaire discriminante sur DCT	Adaptation dépendante des paires, segments critiques. Effets faibles (environ 25% dans le cas idéal)
(Kousidis *et al.* 2008)	CAT	Production de mots cibles + Espace partagé	24 paramètres acoustiques : pitch, intensité, débit de parole, etc.	Adaptation des paramètres en fonction de l'interlocuteur
(Gregory 1986; Gregory *et al.* 1993)	CAT	Interview	Corrélation de pitch	Corrélation plus importante pour les « vraies » interaction (0.77 vs.0.60), le moins élevé socialement converge vers l'autre (r=0.73)

(Flege 1987; Flege and Eefting 1987)	Exemplaires	Perception et Production en Anglais pour des personnes Néerlandaises	VOT de [p], [t] et [k]	VOT croissant avec le niveau des Néerlandais en Anglais
(Bullock *et al.* 2006)	Exemplaires	Production en Anglais/Espagnol pour des personnes bilingues	VOT de [p], [t] et [k]	Influence de la L1 pendant les productions en L2
(Sato *et al.* 2010; Sato *et al.* 2011)	Exemplaires	Production de voyelles + imitation	Pitch, F1	Adaptation significative aux stimuli et plus importante pendant l'imitation
(Gentilucci and Bernardis 2007)	Exemplaires	Production de phonèmes cibles sous différentes conditions (A, V et AV)	Ouverture/fermeture des lèvres, F1, F2, pitch, intensité, durée	Adaptation significative des participants dépendante des conditions (A, V, AV)
(Jones and Munhall 2000; Jones and Munhall 2002)	How/What	Production d'un mot cible + Perturbation environnementale	Pitch	Compensation significative du pitch
Garnier (Garnier *et al.* 2010)	How/What	Espace partagé + Perturbation environnementale	Intensité, pitch, spectre, F1, durée, chute d'intonation, paramètres articulatoires	Augmentation significative des paramètres en présence de bruit
(Aubanel *et al.* 2011)	How/What	Discussion libre + Perturbation environnementale	Intensité, pitch, F1, le débit de parole	Augmentation des paramètres en présence de bruit

Table 1. 1. Tableau récapitulatif des articles présentés dans l'état de l'art

Chapitre 2 Scénario d'interaction pour l'étude de la convergence phonétique

Différents types de scénarios ont été mis en place pour étudier le phénomène de convergence phonétique. Nous allons ici présenter ceux utilisés dans la littérature et celui que nous avons défini en essayant de prendre en compte tous les facteurs qui vont influencer l'intensité du phénomène. L'objectif essentiel qui a gouverné notre choix est de pouvoir collecter un nombre suffisamment de données pour être en mesure d'associer des tests statistiques à nos résultats ainsi que de pouvoir questionner a postériori les facteurs que nous n'aurions pas pu contrôler a priori.

Ce chapitre est organisé de la manière suivante : après une taxonomie des scénarios utilisés dans la littérature et des protocoles utilisés pour collecter les espaces phonétiques de référence, nous présenterons le scénario original mis en œuvre dans le cadre de cette thèse : le jeu des *dominos verbaux*.

2.1 Revue de la littérature

Mis à part les études essayant de dégager des dynamiques d'alignement lors d'interactions libres, le principe général des scénarios utilisés est de collecter des segments phonétiques-cibles – de simples voyelles à des mots en passant par des syllabes – impliqués de manière récurrente dans des productions verbaux. Si de nombreuses recherches ont été effectuées en demandant à des locuteurs de produire des énoncés en réponse directe à des stimuli pré-enregistrés – on parlera alors d'imitation ou de répétition – ou en alternance avec des tâches effectuées par d'autres – on parlera alors de production ambiante –, notre revue s'attachera essentiellement aux scénarios d'interaction réelle.

2.1.1 Pré-tests et post-tests

Les productions hors-interaction de ces segments phonétiques-cibles sont par ailleurs collectés avant ou après la séance d'interaction – lors de séances dites de pré-test et de post-test – afin respectivement de constituer l'espace de référence de chaque locuteur ou de vérifier si les phénomènes observés lors de l'interaction dépendent ou non des stimuli délivrés par l'interlocuteur. Dans le cas où les phénomènes perdurent malgré l'absence de stimuli, on parle de mimesis ou de modification des représentations internes du locuteur utilisées pour planifier ses productions vocales de manière autonome.

Pour ces séances de contrôle, la plupart des études utilisent soit une tâche de lecture des segments-cibles soit une tâche identique à la tâche interactive mais que les sujets exécutent seuls. La première condition peut entraîner le risque de détecter un changement d'élocution ou de style plutôt qu'une véritable adaptation. La deuxième condition, lorsqu'elle peut être appliquée (comme dans la tâche de description d'images dans Delvaux and Soquet, 2007), est nettement plus pertinente et permet de conditionner les sujets à la tâche.

Le pré-test est souvent répété après la tâche pour tester le phénomène d' « after-effect » cité plus haut.

2.1.2 Les interactions avec des stimuli préenregistrés

2.1.2.1 Les tâches d'imitation et de répétition

Les tâches de répétition, voire d'imitation, sont abondamment utilisées dans la littérature car elles permettent aux chercheurs de collecter facilement un grand nombre de productions de segments-cibles tout en contrôlant de manière assez fine – y compris par des techniques de synthèse de parole – la distribution de certains paramètres phonétiques (tels que le contour prosodique, l'intensité moyenne ou le VOT de certains sons). Cela facilite en outre l'analyse des données a posteriori car le contrôle des productions favorise l'utilisation de la segmentation automatique. Cependant ce type de tâche influence grandement le phénomène de convergence.

Sato *et al.* (2011) ont demandé à des sujets de répéter des voyelles isolées, présentées acoustiquement via des haut-parleurs, soit en leur demandant de les répéter soit en les instruisant explicitement d'imiter les stimuli. Cette tâche couplée avec le pré-test et le post-test leur ont permis de collecter environ 14000 exemplaires de voyelles cibles.

Babel (2008; 2009) a utilisé un paradigme similaire mais a ajouté une condition audio-visuelle pour tester si l'attractivité du sujet de référence avait un impact sur la force de la convergence. Les sujets répétaient des mots de fréquence lexicale faible (Goldinger, 1998) contenant les voyelles [i], [æ], [ɑ], [o] et [u]. Cela lui a permis de collecter 400 stimuli (50 en pré-test, 300 pendant la tâche et 50 pendant le post-test) par sujet. Elle a testé 113 sujets dont 53 hommes et 64 femmes.

Gentilucci et Bernardis (2007) a demandé à quatorze femmes (âgées de 22 à 25 ans) de reconnaître et reproduire des chaînes de phonèmes ([aba], [ada] et [aga]) sous différentes conditions. Pendant la première condition, les sujets voyaient le visage d'un acteur prononçant les stimuli mais sans le son. La deuxième condition correspondait à une condition acoustique pendant laquelle les sujets entendaient la voix de l'acteur. Enfin la troisième condition était une condition audiovisuelle (visage et voix de l'acteur). Pendant chaque bloc, 8 [aba], 4 [ada] et 4 [aga] – soit 64 stimuli – ont été présentés. D'après la littérature, les mouvements labiaux des hommes sont plus grands (Elyan 1978; Simpson, 2003) et les formants sont plus bas que ceux des femmes (Hillenbrand *et al.* ,1995): l'auteur s'attendait donc à ce que les mouvements labiaux des femmes s'amplifient pendant la phase de test et que leurs formants s'abaissent.

Jones et Munhall (2000; 2002) ont défini une tâche de répétition pendant laquelle ils ont perturbé le retour auditif de leurs sujets. Dix-huit hommes entre 18 et 30 ans ont dû prononcer le mot « awe » présenté sur un écran d'ordinateur pendant trois secondes. Pendant l'expérience, les sujets étaient enregistrés via un microphone et entendaient en quasi temps-réel leur propre production modifiée ou non via un casque. Trois conditions expérimentales ont été testées. Une condition « shift-up » pendant laquelle on leur renvoyait leur voix dont la fréquence fondamentale était augmentée (l'incrémentation était assez faible pour que le sujet ne se rende pas compte du changement), une deuxième condition « shift-down » où le même procédé a été utilisé mais cette fois en diminuant progressivement la fréquence fondamentale et enfin une dernière condition dite de contrôle pendant laquelle la boucle de

perception n'était pas modifiée. Chaque condition a été divisée en plusieurs blocs : un premier bloc de 10 répétitions pendant lequel la fréquence fondamentale n'était pas modifiée puis un bloc de 100 répétitions pendant lequel la fréquence fondamentale était incrémentée ou diminuée de 1 cent par rapport à la fréquence de référence du sujet, un avant dernier bloc composé de 20 répétitions pendant lequel le sujet entendait son enregistrement augmenté ou diminué de 100 cent et enfin un dernier bloc de 10 répétitions identique au premier bloc. Ils obtenaient donc un total de 140 répétitions pour chaque condition.

2.1.2.2 Les tâches de description

Les tâches de description peuvent être également utilisées pour récolter un corpus de données. Ces tâches permettent de récupérer de nombreuses occurrences de phonèmes tout en contrôlant le contenu du discours de manière implicite. Elles peuvent en outre être remplies par des sujets n'ayant pas accès à la lecture, notamment les enfants. Ce type de tâche a été mise en place par Delvaux et Soquet (2007) et par Cibelli (2009).

Delvaux et Soquet (2007) ont étudié l'influence de la parole ambiante sur les productions de leurs sujets en comparant deux régiolectes différents, le français en Wallonie et le français en Flandre. Ils ont utilisé une tâche de description pour tester si les productions de voyelles caractéristiques des régiolectes des sujets étaient significativement différentes de leurs productions d'origine et significativement proche de celles de l'autre dialecte pendant la tâche. Le sujet était assis face à un écran d'ordinateur, avec une enceinte de chaque côté de l'écran. Le sujet devait alors décrire l'image en prononçant « C'est dans X qu'il y a N Y », X pouvant être {*caisse, frigo*}, N étant {*1, 2, 3*} et Y pouvant être {*stylo, bouquin, tortue, cageot, caisse enregistreuse, fleur, bombe, gourmand, main, mur*} (voir Figure 2. 1). Les sujets testés étaient huit femmes âgées de 22 à 28 ans.

Les auteurs ont pris quelques précautions pour définir leur tâche :

- Un seul sujet de référence a été choisi pour assurer l'homogénéité des stimuli
- La phrase à prononcer a été définie telle que les voyelles étudiées ne soient pas contenue dans le mot en position finale
- Toutes les conditions ont été enregistrées le même jour
- Différentes phases de test ont été enregistrées pour étudier l'évolution de la convergence dans le temps

Les auteurs ont défini trois conditions expérimentales. Pendant la première condition, les sujets décrivaient ce qu'elles voyaient sur l'écran. Les productions des sujets étaient enregistrées, cela correspondait au pré-test pour obtenir un espace phonétique de référence pour chaque sujet. La deuxième condition était ensuite enregistrée. Pendant le « Test », les sujets provenant d'une des deux régions étudiées devaient décrire l'image lorsque qu'une flèche en bas de l'écran pointait vers le bas. Dans les autres cas, elles entendaient la description faite par le sujet de référence parlant l'autre régiolecte. Après la phase de test, les sujets ont de nouveau été enregistrées seules pour observer si elles gardaient des traces de la tâche. Les enregistrements ont permis de recueillir 200 occurrences de « caisse » et 200 de « frigo » par sujet par condition. Cibelli (2009) a utilisé une tâche similaire mais a étudié l'adaptation entre deux langues différentes.

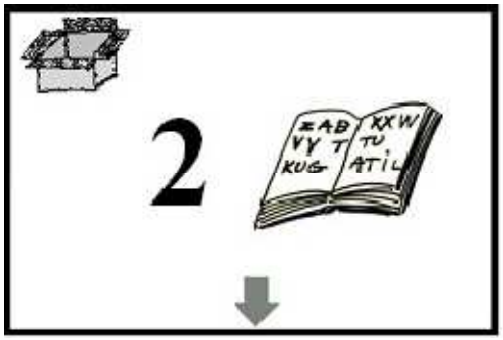


Figure 2. 1. Écran d'ordinateur présenté aux sujets pour procéder à la tâche de description de Delvaux et Soquet (2007). Ils doivent alors prononcer : « C'est dans la caisse qu'il y a deux bouquins ». La flèche vers le bas signifie que c'est au tour du sujet de parler. Les analyses ont été faites sur deux segments critiques dont le /ɛ/ contenu dans /kɛs/.

Le problème avec les tâches de description est qu'elles ne correspondent pas à une réelle situation d'interaction, on peut considérer que les sujets ne développent pas de stratégies de communication au cours de la tâche. Elles ne permettent donc pas d'étudier le phénomène de convergence phonétique. De plus dans les deux cas présentés, les analyses ont été faites sur quelques segments critiques caractéristiques des différents régiolectes ou langues et ne fournissent pas une vision plus globale/structurale de la transformation de l'espace acoustique.

2.1.3 Les espaces partagés

De nombreux chercheurs ont utilisé des tâches de collaboration impliquant des espaces partagés pour étudier la convergence phonétique en interaction. Comme les sujets doivent effectuer la tâche en équipe (souvent en duo), ils adoptent différentes stratégies pour que leur interaction soit la plus efficace possible. Ces tâches en équipe permettent aussi d'étudier le lien entre l'amplitude de l'adaptation et le rôle des sujets pendant l'interaction.

Kim *et al*. (2011) ont utilisé un jeu pendant lequel les sujets devaient trouver 10 différences entre les images qu'ils avaient devant eux (voir Figure 2. 2). Cela les a obligé à prononcer plusieurs fois le nom d'icônes présentes sur les images. 16 conversations ont été enregistrées et 192 mots (16 conversations x 3 mots x 2 moments de l'interaction x 2 locuteurs) ont été extraits de ces conversations pour mettre en place un test XAB afin de souligner le phénomène d'adaptation.

Pardo (2006) a utilisé un concept développé au Human Communication Research Center de l'Université de Glasgow et Edinbourg nommée « Map Task » (Anderson *et al.* ,1991). Cette tâche utilisait deux cartes contenant les mêmes illustrations (voir Figure 2. 3). Une des deux cartes contenait un chemin spécifique qui passait autour de certaines illustrations avec un point de départ et un point d'arrivée. La deuxième carte contenait seulement le point de départ et le point d'arrivée. Les sujets devaient alors collaborer pour que le sujet qui voyait la carte aide le deuxième sujet à la reproduire correctement. Cette tâche a permis de prendre en compte plusieurs facteurs. L'impact du rôle social a pu être testé car il y avait une hiérarchie établie entre les deux sujets (celui qui donnait les instructions vs celui qui recevait les instructions). Des paires de même sexe ont effectué cette tâche pour éviter le phénomène de dominance sociale.

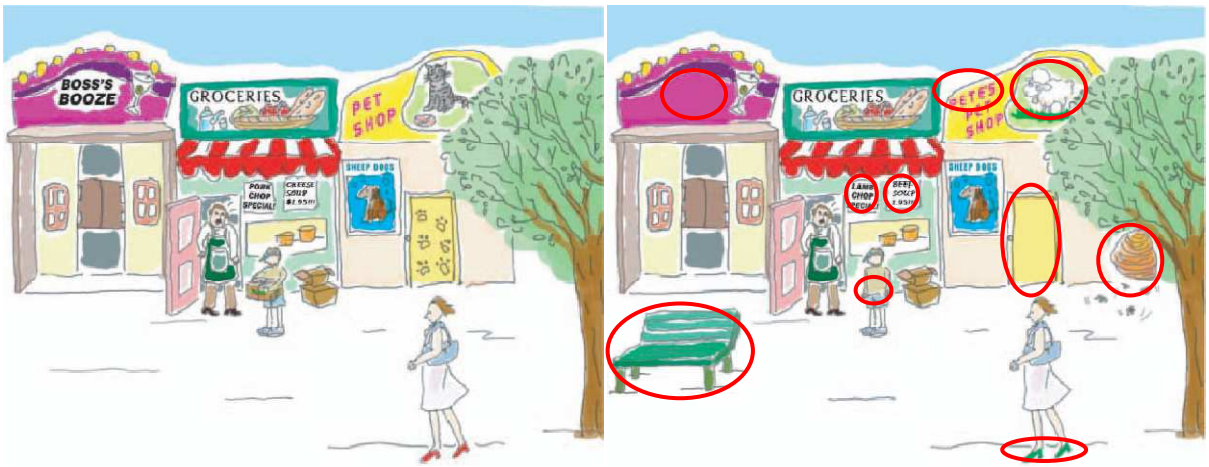


Figure 2. 2. Jeu des 10 différences utilisé par Kim *et al*. (2011).

Les productions des sujets ont été enregistrées avant (pré-test) et après (post-test) l'interaction pour obtenir une référence pour chaque sujet et également testé le phénomène d' « after-effect ». Pendant le pré-test et le post-test les sujets ont répété le nom des illustrations qui se trouvaient sur les cartes de façon à obtenir une prononciation de référence des mots utilisés pendant l'interaction. Les enregistrements des pré-tests ont également été utilisés pour appareiller les sujets, les auteurs ont choisi de créer des paires dont les f0 étaient approximativement les mêmes. Les pré-tests ont été enregistrés une à deux semaines avant l'interaction et les post-tests ont été enregistrés juste après l'interaction.

Pendant l'interaction, chaque rôle (donneur/receveur) a été assigné puis les sujets ont dû reproduire cinq cartes différentes. Ils étaient séparés par une cloison pour éviter que les cartes ne soient visibles à l'autre. Six hommes et six femmes ont participé à cette expérience. Pardo a également utilisé un test de perception pour mettre en évidence la convergence phonétique. 24 paires de mots ont été extraits des différentes conditions d'enregistrements (pré-test, post-test, début et fin d'interaction, donneur répété par receveur et receveur répété par donneur) pour créer un test AXB.

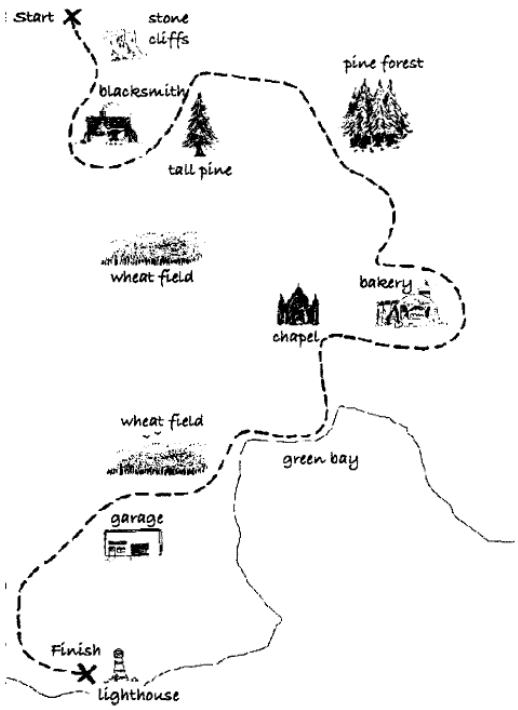


Figure 2. 3. Map task utilisé par Pardo (2006)

Aubanel *et al.* (Aubanel and Nguyen, 2010; Aubanel, 2011) ont crée un scénario qu'ils ont appelé GMUP (pour Group 'em up) pendant lequel 24 sujets regroupés par paire de même sexe ont dû procéder à une tâche collaborative. Cette tâche était divisée en deux étapes :

- Pendant une première phase, les sujets ont dû associer 16 noms de famille (qui ont été inventés par les auteurs afin de récupérer le matériel phonétique nécessaire pour leur analyse) à 16 photos qui se trouvaient sur un écran d'ordinateur face à chaque sujet. Ainsi ils devaient prononcer les noms plusieurs fois en décrivant chaque photo.
- Lors de la seconde phase, les sujets ont dû diviser les seize personnages en quatre groupes de taille variable en fonction d'une caractéristique donnée pour chacun d'entre eux. Les noms des groupes n'étaient pas imposés.

Les seize noms de famille ont été inventés par les auteurs afin de récupérer le matériel phonétique nécessaire pour leurs analyses. Les sujets ont d'abord répondu à deux questionnaires. Le premier (basé sur le PFC pour *Phonologie du Français Contemporain* de *Durand et al., 2003*) avait pour but de déterminer la variété régionale du Français parlé par chaque sujet. Pour le deuxième questionnaire, Aubanel et Nguyen se sont inspirés de l'échelle de Désirabilité Sociale de Crowne et Marlowe *(*Crowne and Marlowe, 1960; Marlowe and Crowne, 1961*)* qui traduit l'envie de chaque sujet d'être accepté socialement pendant une interaction (son score sera alors plus élevé sur une échelle de 1 à 10).

Après ces deux questionnaires, les auteurs ont sélectionné 24 sujets (20 filles et 4 garçons, d'âge moyen 15,8 ans). Ils ont été choisis de façon à avoir 12 locuteurs du Français standard et 12 locuteurs du Français méridional. Ils ont alors pu être regroupés par paire de même sexe et de score similaire sur l'échelle de Désirabilité Sociale. Une semaine avant l'expérience, les sujets ont été enregistrés prononçant 3 fois les 16 noms de famille contenus dans des phrases (pré-test). Cet enregistrement a permis d'obtenir une prononciation de référence pour chaque sujet. Chaque paire a ensuite accompli

le jeu à trois reprises ce qui a permis de récupérer en moyenne 4.5 répétitions de chaque noms. Enfin chaque sujet a répété la phase du pré-test juste après les trois jeux pour tester le phénomène d' « after-effect ».

Figure 2. 4. Jeu de survie utilisé par Kousidis *et al*. (2009).

Kousidis *et al.* (2009) ont étudié le phénomène d'adaptation grâce à un scénario durant lequel cinq paires de sujets devaient collaborer dans un jeu de survie. Ils se trouvaient dans trois situations possibles : ils pouvaient être naufragés, ou coincés dans une navette spatiale ou bien encore perdus dans les montagnes neigeuses de l'Himalaya. Ils avaient alors 10 minutes – le temps pour que les secours arrivent – pour classer 15 items, qui leur étaient présentés sur un écran par des imagettes (voir Figure 2. 4), par ordre d'importance pour leur survie (1 pour le plus important et 15 pour le moins important). Pas d'analyse dédiée aux mots-clés n'a cependant été décrite.

Dans le cadre de l'ANR AMORCES, nous avons mis en place une tâche de collaboration pour étudier l'apport du regard dans l'accomplissement d'une tâche et comparer le comportement d'un homme face à un autre humain ou à un robot (Fagel *et al.* ,2010). Deux sujets étaient assis face-à-face à une table, le plateau de jeu étant au centre de la table (voir Figure 2. 5). Chacun avait un rôle bien précis pendant l'interaction. Le manipulateur entendait un label parmi {P, T, B, D, G, M, N, F, S}. Ne sachant pas où se trouvait le cube correspondant à ce label, il demandait à l'informateur – en prononçant le label – de lui indiquer sa position grâce à une couleur {rouge, bleu, vert} et une lettre {A, I O}. Le manipulateur déplaçait alors le bon cube sur l'emplacement opposé à celui d'origine. Nous avons comparé les cas où l'informateur portait ou non des lunettes noires – grâce à un eyetracker – pour démontrer que le regard était un élément essentiel pendant une interaction face-à-face.

Figure 2. 5. Jeu de cubes utilisé par Fagel *et al.* (2010).

Le problème de ces scénarios impliquant la fonction déictique est qu'ils ne permettent pas de récupérer un grand nombre d'échantillons de parole, celle-ci étant parfois complémentaire à d'autres modalités. De plus, comme pour les tâches de description, les analyses ne sont faites que sur les segments critiques et pas sur une couverture plus exhaustive de l'espace acoustique du locuteur.

2.1.4 La parole libre

Pour étudier le phénomène de convergence phonétique, des chercheurs ont choisi d'enregistrer leurs sujets en parole libre. Gregory *et al.* (1996) ont étudié 25 conversations tirées d'interviews télévisées de Larry King. Kousidis *et al.* (2008) ont étudié les conversations d'un sujet de référence avec 3 sujets testés récupérant ainsi 83.7 minutes de parole. Lee *et al.* (2010) ont utilisé l'enregistrement de couples mariés pendant une séance de thérapie pour récupérer leur corpus. Certaines bases de données sont souvent utilisées : citons notamment le « Columbia Games Corpus » développé par Benus (2009) pendant lequel des paires doivent collaborer pour résoudre des jeux sur ordinateur (Nenkova *et al.* ,2008; Heldner *et al.* ,2010; Levitan and Hirschberg, 2011). Enfin, De Looze *et al.* (2011) ont utilisé le corpus « D64 » enregistré par Oertel *et al*.(2010) pour étudier la convergence en conversation spontanée et le lien entre l'amplitude de la convergence et le degré d'implication du locuteur.

C'est la situation la plus écologique qui puisse être utilisée mais c'est également celle qui va entraîner le plus de contraintes au niveau des analyses. Comme les données ne sont pas contrôlées, toutes les annotations vont devoir être faites manuellement (comme pour les 9 heures et 8 minutes du « Columbia Games Corpus », Levitan & Hirschberg, 2011).

2.2 Les dominos verbaux

Les études précédentes ont mis en relief différentes contraintes obligatoires pour étudier le phénomène de convergence phonétique dans toute sa complexité :

- La tâche doit impliquer une collaboration des sujets
- La tâche doit permettre de récupérer un grand nombre d'exemplaires

- La tâche doit permettre de contrôler les données prononcées

Pour remplir toutes ces conditions, nous avons choisi d'utiliser un jeu de langage appelé « Dominos verbaux » (Bailly and Lelong, 2010; Lelong and Bailly, 2011). Ce jeu est bien connu des cours de récréation et favorise l'apprentissage de la segmentation des mots en syllabes (Arléo, 1997).

Le principe du jeu est simple. Les deux sujets se trouvent face-à-face, l'un des deux sujets prononce un mot – par exemple « rotor » – le deuxième sujet aura alors deux choix possibles – « tordu » ou « berlue » –, il devra alors choisir le mot qui commence par la syllabe finale de « rotor » soit « tordu » et ainsi de suite (voir Figure 2. 6).

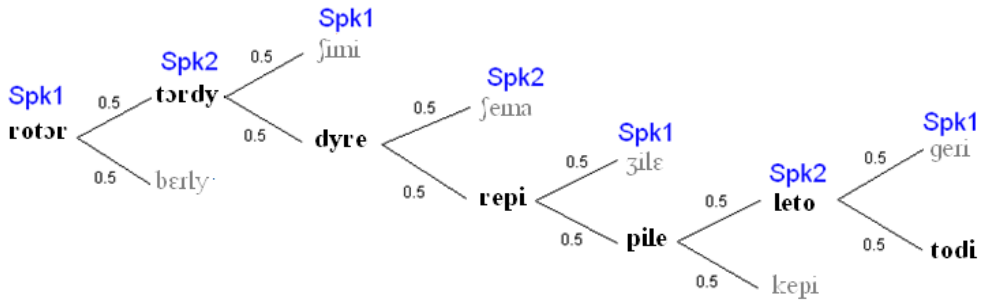


Figure 2. 6. Succession des premiers dominos verbaux, les solutions sont mises en évidence en gras

Nous avons opté pour des mots dissyllabiques tels que **bateau** [bato], **taudis** [todi], **diffus** [dify], **furie** [fyri], etc. pour que la charge cognitive ne soit pas trop élevée, favorisant ainsi la succession des différents enregistrements. Les mots ont été sélectionnés de façon à collecter uniformément les variations allophoniques des 8 voyelles orales périphériques du Français: [a], [ɛ], [e], [i], [y], [u], [o], [ɔ].

La chaîne de dominos a été construite de manière à forcer l'attention mutuelle. Un sujet doit forcément attendre que son partenaire ait prononcé le mot pour pouvoir faire son choix pour le tour suivant. En effet, lorsque le locuteur 1 a prononcé « rotor », son écran indique les mots « chimie » et « durée », comme les mots « torchis » et « tordu » – que le locuteur 2 peut choisir de prononcer – existent tous les deux en Français, il doit attendre le tour du locuteur 2 – qui prononce effectivement « tordu » s'il choisit le bon domino – pour choisir le prochain mot à prononcer.

Les expériences ont été menées sur deux corpus différents. En première approche, nous avons utilisé un corpus contenant 190 mots, soit 95 mots prononcés en interaction par chaque sujet. En appliquant la reconnaissance de parole, nous avons remarqué que le corpus n'était pas assez important. Pour cela, nous avons utilisé la moitié des données du pré-test pour créer un modèle HMM de chaque locuteur et nous avons calculé le score de reconnaissance sur l'autre moitié du pré-test. Nous avons alors obtenu des scores plutôt faibles, nous avons donc augmenté la taille du corpus pour les expériences suivantes en utilisant 175 mots prononcés en interaction par chaque locuteur. Cette chaîne de dominos nous a alors permis de récupérer environ 40 exemplaires de chaque voyelle (voir Table 2. 1).

phonèmes	a	ɛ	e	i	y	u	o	ɔ	ø,œ
#items corpus I	26	29	22	22	24	21	20	16	6
#items corpus II	47	48	45	43	44	40	43	31	9

Table 2. 1. Nombre de réalisations des phonèmes collectés pour chaque interlocuteur pendant le jeu pour chaque corpus.

2.2.1 Conditions d'interaction

Les dominos ont été prononcés sous différentes conditions. Comme pour les études précédentes, il a d'abord fallu obtenir des prononciations de référence pour chaque sujet. Nous avons donc défini une première condition appelée « **Pré-test** » pendant laquelle les sujets doivent lire les 350 dominos qui seront prononcés pendant la phase de jeu par les deux partenaires. Cela permet de caractériser les espaces phonétiques de chaque sujet et de mesurer l'amplitude de la convergence s'il y en a.

Juste après l'enregistrement de leur pré-test, les sujets ont joué au jeu de dominos. Ils ont donc prononcé chacun 175 mots. L'enregistrement des pré-tests nous permet également d'ajouter une condition de perturbation **ambiante** à notre paradigme pendant laquelle les sujets interagissent avec l'enregistrement de leur partenaire sans le savoir.

Nous pouvons également tester si la connaissance du contenu linguistique à prononcer va influencer les productions des sujets en leur demandant, dans un premier de temps, de **répéter** le mot précédemment énoncé par leur partenaire avant de faire leur choix. Nous pourrons alors comparer les taux de convergence des mots répétés et des mots de l'interaction.

Enfin, après chaque enregistrement, les sujets ont lu de nouveau les 350 mots prononcés pendant l'interaction. Cette condition « **Post-test** » sert à caractériser le phénomène d' « after-effect » pour savoir si l'interaction a laissé des traces qui ont perturbées à plus ou moins long-terme les modèles internes des sujets.

2.2.2 Conditions d'enregistrement

Nous avons mis en place quatre types d'expériences différents. Dans un premier temps (**Expérience I**), nous avons demandé à de parfaits inconnus de procéder à la tâche. Ils ne s'étaient jamais rencontrés et ont interagi de manière médiatisée. Cela a été possible grâce à la plateforme MICAL qui est composée de deux salles séparées par un miroir tinté. Chaque sujet se trouvait dans une salle et ils communiquaient entre eux grâce à des écouteurs et des microphones.

Dans un deuxième temps (**Expérience II**), nous avons demandé à des collègues qui se connaissaient de longue date de jouer au jeu de dominos, toujours en face-à-face médiatisée (ils se connaissent en moyenne depuis 15 ans, de 10 à 25 ans).

Dans un troisième temps (**Expérience III**), des amis (moyenne de 2 ans et 9 mois, de 6 mois à 8 ans) ont participé au jeu en face-à-face (voir Figure 2. 7).

Figure 2. 7. Interaction face-à-face de l'expérience III. Pendant cette expérience les mouvements de têtes ont été enregistrés grâce au système de capture de mouvement Qualysis.

Enfin, pour la dernière expérience (**Expérience IV**), nous avons fait interagir des personnes d'une même famille et nous leurs avons demandé de répéter le mot de leur partenaire avant de prononcer leur mot.

Pour toutes les expériences présentées, nous avons demandé aux sujets d'éviter de parler en même temps et également de prononcer de nouveau le mot s'ils s'étaient trompés. Ces précautions ont été prises afin de faciliter la segmentation automatique et l'alignement.

2.2.3 Paramètres expérimentaux

Pour les expériences I et II, les locuteurs communiquaient grâce à des microphones et des écouteurs. Les signaux ont été enregistrés grâce à une carte son de bonne qualité à une fréquence d'échantillonnage de 16 kHz. Les dominos étaient présentés sur un écran d'ordinateur en version pdf.

Pour les expériences III et IV, les locuteurs étaient assis face-à-face, à une table, avec deux écrans d'ordinateur face à eux. Ils ont été enregistrés grâce à une caméra, un miroir nous permettant de voir les deux participants (voir Figure 2. 7). Nous avons utilisé des claviers connectés aux ordinateurs pour que les participants puissent progresser dans le jeu : lorsque le sujet avait prononcé son domino, il appuyait sur une touche du clavier pour faire apparaître les deux dominos suivant sur son écran.

2.2.4 Corpus et participants

44 personnes – 18 hommes et 26 femmes – âgées en moyenne de 28 ans et 7 mois (médiane de 25 ans, minimum 18 ans et maximum de 53 ans), ont formé 35 paires pour participer aux diverses expériences. 15 personnes, 8 hommes et 7 femmes qui ont composé 12 paires ont été enregistré sur le premier corpus. Ce corpus nous a permis de récupérer 1 heure 41 de parole pendant le pré-test (moyenne : 4 min 46, déviation standard : 51 sec) ainsi qu'1 heure 41 de parole en interaction (moyenne : 4 min 11, déviation standard : 27 sec). Nous utilisons le deuxième corpus sur deux expériences différentes. 17 personnes – 5 hommes et 12 femmes – ont été répartis en 13 paires pour participer à la première expérience avec le second corpus. 2 heures 22 de parole ont été enregistrées pendant le pré-test (moyenne : 7 min 55, déviation standard : 1 min 46) et 3 heures 53 en interaction (moyenne : 8 min 59, déviation standard : 1 min 40). Enfin, 12 personnes – 5 hommes et 7 femmes – divisés en 10 paires, ont joué au jeu de dominos avec les répétitions. 1 heure 32 de parole ont été récupéré pendant le pré-test (moyenne : 7 min 40, déviation standard : 1 min 27) et 4h02 en

interaction (moyenne : 12 min 05, déviation standard : 1 min 12) (voir Table 2. 2). Il faut également noter que quatre locuteurs ont été plusieurs fois sujet de référence pour plusieurs interactions, ce qui nous permettra d'observer si le comportement d'adaptation dépend de l'interlocuteur. Ils ont été mis en évidence à l'aide d'un code couleur dans Table 2. 3. Les quatre locuteurs sont un homme de 50 ans (en condition « inconnus » et « amis »), une femme qui a interagi à plusieurs stades des expériences et donc à des âges différents (i.e. 24, 25 et 26 ans en condition « inconnus », « amis » et « famille »), un homme de 21 ans (en condition « inconnus ») et un homme de 25 ans (en condition « famille »).

Les corpus nous permettent également d'étudier l'influence de la fréquence lexicale sur l'amplitude de l'interaction. On suppose que plus la fréquence lexicale d'un mot est faible, moins les sujets possèdent de modèles internes de ce mot, plus ils vont avoir tendance à adopter une stratégie orientée vers l'auditeur donc converger avec plus de force. Les distributions des fréquences lexicales des divers mots des deux chaînes de dominos – calculées à partir des fréquences lexicales de la base de données LEXIQUE (http://www.lexique.org/) – sont données à la Figure 2. 8 pour les corpus I et II.

	Condition	Total	Moyenne	Déviation Standard
Corpus 1	Pré-test	1h41	4min46	51 s
	Interaction	1h41	4min11	27 s
Corpus 2	Pré-test	2h22	7min55	1min46
	Interaction	3h53	8min59	1min40
	Pré-test	1h32	7min40	1min27
	Interaction avec Répétition	4h02	12min05	1min12

Table 2. 2. Corpus récupérés grâce aux « Dominos Verbaux »

Le paradigme proposé permet non seulement de prendre en compte toutes les contraintes imposées pour étudier le phénomène de convergence phonétique mais aussi de comparer nos résultats à ceux des études précédentes (répétitions, perturbation ambiante, test AXB, etc.). Il facilite la segmentation automatique en contrôlant les mots qui vont être prononcés au cours de l'interaction tout en forçant la collaboration entre les sujets (voir Table 2. 3).

Paire	Corpus	Conditions	Sujet 1		Sujet 2	
			Sexe	Age	Sexe	Age
1	Court	Inconnus	homme	50	femme	20
2	Court	Inconnus	homme	50	femme	22
3	Court	Inconnus	homme	50	femme	25
4	Court	Inconnus	homme	50	homme	21
5	Court	Inconnus	homme	50	homme	22
6	Court	Inconnus	homme	50	homme	23
7	Court	Inconnus	femme	24	femme	21
8	Court	Inconnus	femme	24	homme	20
9	Court	Inconnus	femme	24	femme	21
10	Court	Inconnus	homme	21	homme	29
11	Court	Inconnus	homme	21	homme	22
12	Court	Inconnus	homme	21	femme	25
13	Long	Amis	homme	50	homme	50
14	Long	Amis	homme	50	homme	39
15	Long	Amis	homme	50	homme	39
16	Long	Amis	femme	25	homme	51
17	Long	Amis	femme	25	homme	26
18	Long	Amis	femme	25	homme	24
19	Long	Amis	femme	25	femme	24
20	Long	Amis	femme	25	femme	25
21	Long	Amis	femme	18	femme	18
22	Long	Amis	femme	21	femme	21
23	Long	Amis	femme	26	femme	25
24	Long	Amis	femme	25	femme	23
25	Long	Amis	femme	25	femme	22
26	Long	Amis	femme	26	homme	27
27	Long	Amis	femme	26	femme	26
28	Long	Famille	femme	26	femme	23
29	Long	Famille	femme	26	femme	31
30	Long	Famille	femme	26	femme	51
31	Long	Famille	femme	26	homme	53
32	Long	Famille	homme	25	homme	50
33	Long	Famille	homme	25	femme	49
34	Long	Famille	homme	25	homme	23
35	Long	Famille	homme	25	femme	19

Table 2. 3. Tableau récapitulatif des locuteurs, des conditions utilisées et des corpus. Chaque couleur correspond à un des quatre locuteurs qui ont interagi avec plusieurs personnes.

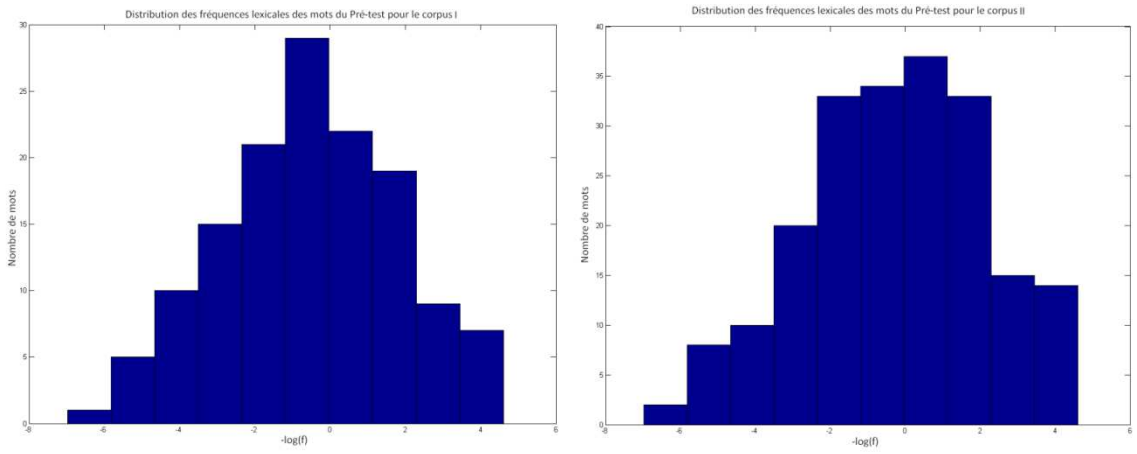


Figure 2. 8. Distribution des fréquences lexicales des mots du Pré-test pour les corpus I et II.

2.3 Commentaires

Nous avons proposé ici un paradigme original de recueil d'items lexicaux, que nous nommerons par la suite « dominos verbaux ». Ce paradigme nous permet d'enregistrer nos sujets sous plusieurs conditions. Nous pouvons faire prononcer les différents items aux sujets pendant un pré-test afin de récupérer un espace phonétique de référence utile pour caractériser l'amplitude de la convergence. Nous aurions également pu demander aux sujets de jouer seuls afin d'obtenir de meilleurs conditions de comparaison au niveau de l'énonciation (lecture vs. interaction). Cette étape peut être répétée après l'interaction afin de tester le phénomène d' « after-effect » c'est-à-dire tester si l'interaction à laisser des traces qui vont perturber les productions des sujets. Comme les sujets sont tous enregistrés en pré-test avant toute interaction, nous pouvons utiliser ces enregistrements afin de mettre en place une phase de jeu en condition ambiante. On pourrait alors analyser le comportement de nos sujets lorsque leurs interlocuteurs ne s'adaptent pas à eux. Enfin, nous pouvons également ajouter une condition de répétition des dominos prononcés par les locuteurs afin de comparer nos résultats en interaction avec ceux obtenus en imitation, qu'elle soit volontaire ou non.

Le paradigme nous permet également de contrôler le nombre d'exemplaires de chaque phonème qui va être récupéré pour nos analyses et aussi de faire varier la fréquence lexicale des mots afin de comparer l'amplitude de la convergence pour les mots de fréquence lexicale faible ou forte.

Il présente cependant quelques inconvénients. Le déroulement du jeu n'est pas très adapté à l'analyse de la convergence des paramètres prosodiques. En effet, les mots choisis étant disyllabiques, il est difficile d'analyser un effet de prosodie sur des énoncés aussi courts. De plus, le paradigme peut induire un effet de liste aussi bien pendant la lecture du pré-test que pendant l'interaction au cours de laquelle un rythme peut se mettre en place. Enfin, les phonèmes que nous allons étudier ne se trouvent pas tous à la même position dans les mots (initiale vs. finale), ce qui peut introduire un problème dû à l'accentuation.

Dans le chapitre suivant, nous allons présenter différentes méthodes pour caractériser la convergence phonétique de manière objective puis les résultats obtenus à partir des corpus

enregistrés avec les différentes expériences de « dominos verbaux » afin d'étudier les différents facteurs qui vont influencer l'amplitude de la convergence phonétique en interaction face-à-face.

Chapitre 3 Caractérisation de la convergence phonologique

A l'inverse de Delvaux et Soquet (2007) et Aubanel *et al.* (Aubanel and Nguyen, 2010; Aubanel, 2011) qui ont centré leur analyse sur des segments sonores porteurs de l'identité dialectale des interlocuteurs, notre corpus n'était initialement pas prévu pour étudier la convergence phonologique. Les sujets venant de différentes régions de la France, nous avons toutefois étudié ce phénomène.

Nous avons observé quelques variations dialectales, particulièrement sur les variations allophoniques des voyelles moyennes. La plupart des participants venaient du nord de la France : ils prononçaient ainsi les voyelles ouvertes exclusivement à l'intérieur de syllabes fermées (i.e. *sabord* [sabɔʀ] vs. *sabot* [sabo]). Certains locuteurs avaient cependant tendance à maintenir une opposition de timbre marquée sur des items lexicaux différenciés orthographiquement (Boula de Mareüil *et al.,* 2010) tels que dans *vallée* et *valais* ([vale] vs. [valɛ]), *miné* et *minet* ([mine] vs. [minɛ]), etc. Ces variantes de prononciation reflètent non seulement l'origine géolinguistique des locuteurs – on peut distinguer trois macro-variétés régionales du français : le français du nord, celui du sud et celui de l'est (Woehrling et Mareüil, 2006; Woehrling *et al.,* 2009) – mais aussi leur volonté inconsciente de montrer leur compétence linguistique et notamment orthographique, cette dernière étant particulièrement valorisée par le protocole d'échange de mots orthographiés utilisé dans notre travail.

3.1 Identification des variantes par reconnaissance de Parole

Nous avons créé des modèles HMM de chaque interlocuteur à partir d'un étiquetage semi-automatique du pré-test. Une explication plus précise sur la reconnaissance de parole sera donnée à la partie 4.3. Nous avons ensuite utilisé ces modèles pour aligner un treillis phonologique incluant pour chaque mot l'ensemble des variantes phonologiques possibles et effectivement observées dans l'ensemble du corpus au fur et à mesure de sa constitution. Ainsi le treillis du mot « lady » dont nous avions anticipé la prononciation ainsi proposée par le Robert : [lɛdi] compte finalement les variantes suivantes : [l̪(ɛ|e)[j]di] avec une ouverture de la première voyelle au choix suivie d'une semi-voyelle optionnelle, typique d'une prononciation à l' « anglaise ».

3.2 Étiquetage manuel

Nous avons ensuite vérifié l'ensemble des étiquettes (voyelles ouvertes fermées ou ouvertes, consonnes dévoisées ou non) à l'aide de Praat (Boersma & Weenink, 2005), en nous basant sur la répartition propre au locuteur de l'espace de réalisation du premier formant (*typiquement F1=500 Hz pour [o] et [e] et F1=700 Hz pour [ɔ] et [ɛ]*). En effet, comme souligné par Neagu (1998) et confirmé par Ménard *et al.* (2008), les locuteurs répartissent de manière relativement consistante les hauteurs moyennes des voyelles mi-ouvertes et mi-fermées – soit de la série [ɔ], [œ], [ɛ] vs. la série [o], [ø], [e]) – avec des valeurs moyennes de F1 très semblables pour chaque voyelle de chaque série mais pas forcément égales à 2/3 vs. 1/3 de l'écart entre le F1 moyen du [a] et celui de la série fermée [u], [y], [i].

3.3 Résultats

Dans la plupart des cas, les différences entre l'étiquetage automatique et manuel restent très faibles. Nous devons cependant signaler que l'étiquetage des variations allophoniques des voyelles moyennes reste très complexe car les locuteurs ont de plus en plus tendance à antérioriser les voyelles moyennes. L'étiquetage est particulièrement difficile en position non accentuée où la coarticulation peut perturber la perception.

Nous avons comparé la proportion de voyelles fermées prononcée par chaque sujet de chaque paire sur le même corpus. Nous avons observé quelques cas d'adaptation phonologique (voir Figure 3. 2 et Figure 3. 3) – i.e. les sujets adoptent une manière de prononcer les voyelles moyennes différente de celle de leur pré-test pour se rapprocher de la prononciation de leur interlocuteur –. Ces adaptations n'étaient cependant pas significatives dû au nombre de données limitées. Si on observe le comportement d'un sujet de référence avec différents interlocuteurs, on remarque que l'adaptation dépend bien de l'interlocuteur (Voir Figure 3. 1). Par exemple, ALa prononce moins de [o] fermé en interaction avec MMP pour s'adapter à lui mais ne le fait pas pendant son interaction avec MSM.

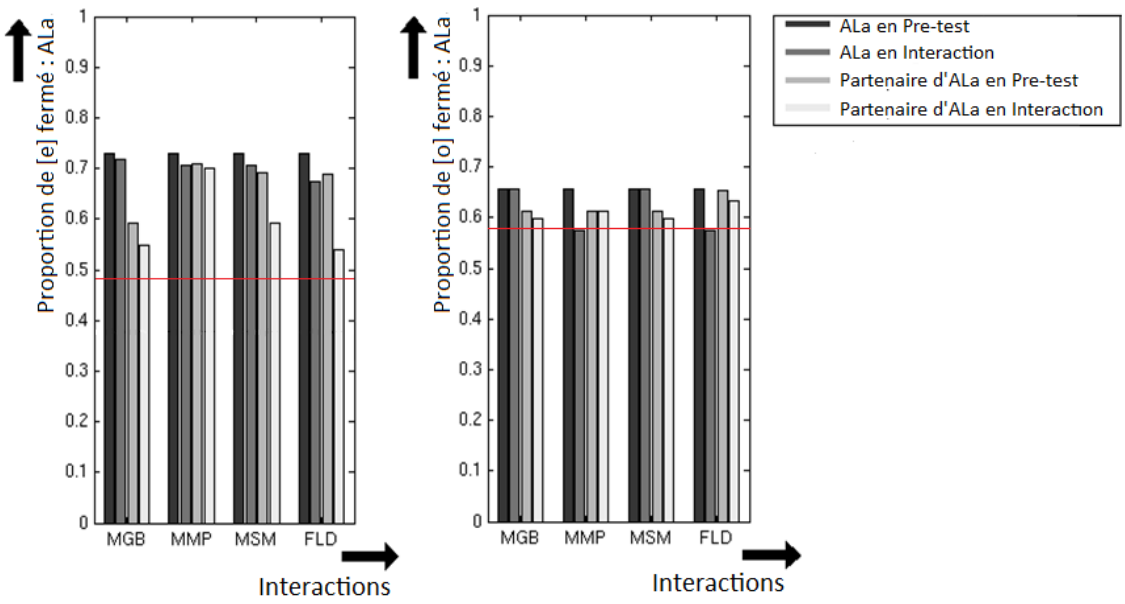


Figure 3. 1. Proportion de voyelles moyennes fermées pour 4 paires ([e] à gauche et [o] à droite). Le sujet de référence est la même femme ALa en interaction avec 3 hommes (MGB, MMP, MSM) et une femme (FLD). Pour chaque paire, les barres représentent la proportion de voyelles fermées prononcées par ALa pendant le pré-test, puis par ALa en interaction avec son interlocuteur, par son interlocuteur en interaction avec elle et enfin pendant le pré-test de son interlocuteur. Comme précédemment, la ligne horizontale rouge correspond à la proportion de voyelles fermées initialement attendue dans le corpus.

Sur la Figure 3. 2, on observe la différence entre la proportion de voyelles moyennes fermées prononcée par les sujets et un choix aléatoire à 50%. Ainsi, si les barres descendent vers le bas, cela signifie que le sujet à tendance à prononcer plus de voyelles moyennes fermées. Nous avons pris soin de comparer les mêmes données. Nous appelons « jeu 1 » les mots prononcé en interaction par le

sujet de référence et « jeu 2 » celui prononcé par le sujet testé. Comme les sujets prononcent, pendant leur pré-test, tous les mots contenus dans l'interaction, nous avons pu comparer ces proportions sur les mêmes corpus. Pour le sujet de référence, nous observons la proportion de voyelles fermées prononcées sur les mots du « jeu 1 » et pour le sujet testé nous observons celle prononcé sur les mots du « jeu 2 ». Sur la Figure 3. 2, la première barre correspond à la proportion de voyelles fermées prononcée par un sujet pendant son pré-test, la deuxième barre correspond à celle prononcée en interaction et la troisième barre à celle de son partenaire pendant son pré-test. Comme pour chaque paire, nous observons ces proportions soit sur le « jeu 1 » soit sur le « jeu 2 », nous n'avons pas de symétrie entre la troisième barre de la ligne du haut et la première barre de la ligne du bas, pour chaque interaction. On remarque que la plupart des sujets sont dans ce cas. Pour chaque sujet de chaque paire (sujet de référence en haut, sujet testé en bas), la première barre correspond à la proportion de voyelles moyennes fermées prononcées pendant son pré-test, la barre du milieu traduit la proportion de voyelles fermées prononcées en interaction et la dernière barre correspond à la proportion de voyelles moyennes fermées prononcées par son partenaire pendant son pré-test. Dans chaque cas, on prend soin de comparer les mêmes ensembles de mots.

Quelques cas de divergence phonologique ou d'absence de convergence ont été mis en évidence en orange pour les femmes et en bleu ciel pour les hommes sur les Figure 3. 2 et Figure 3. 3. Ils sont également mis en évidence respectivement par un « d » ou un « _ ». Les autres cas correspondent à des cas d'adaptation phonologique, un « c » indique si on observe une convergence supérieure à 10%. Ils correspondent au cas où la proportion de voyelles fermées d'un sujet en interaction est comprise entre celle du sujet pendant son pré-test et celle de son partenaire pendant son pré-test.

On observe différents comportements sur la Figure 3. 3 comme une adaptation mutuelle pour les paires 1, 10, 21, 26, une adaptation du sujet de référence vers le sujet testé comme pour les paires 3, 5, 6, 12, 15, 17, 19, 20, 24, une adaptation du sujet testé vers le sujet de référence comme pour les paires 9, 11, 16, 18, et 34. Il existe également quelques cas de divergence comme pour les paires 2, 13, 14 et 23. Il est intéressant de remarquer que sur la dernière partie des figures qui correspondent à l'expérience entre des personnes d'une même famille (sauf pour les paires 26 et 27 qui servent de contrôle), la distance phonologique (différence entre les deux pré-tests) est non seulement faible – la plupart des sujets ne change pas de plus de 10% leur façon de prononcer les voyelles fermées puisqu'ils partagent avec leur partenaire le même système phonologique – mais que seules des convergences sont observées (cf. Figure 3. 3).

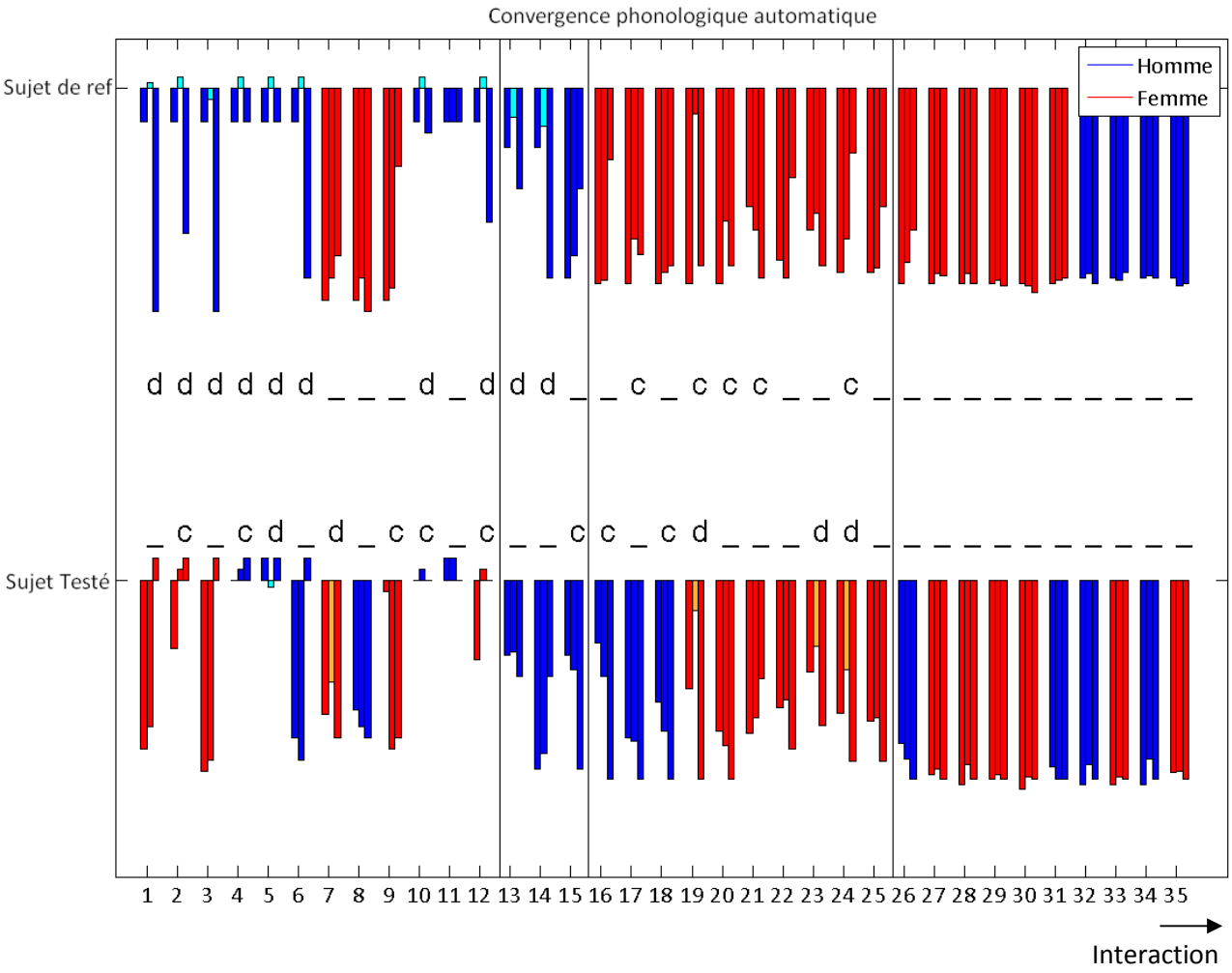


Figure 3. 2. Proportions de voyelles fermées prononcées par les sujets pendant leur pré-test (première barre), l'interaction (deuxième barre) et le pré-test de leur partenaire (troisième barre) pour 35 interactions. Ces proportions ont été calculées en utilisant la reconnaissance de parole. Les barres indiquées en bleu ciel ou en orange correspondent à des cas de divergence phonologique. Les étoiles correspondent à des convergences phonologiques supérieures à 10%.

Nous avons calculé la corrélation entre les proportions trouvées par chaque méthode pour chaque condition et pour chaque corpus. Ils sont résumés dans les Table 3. 1 et Table 3. 2. Les différences les plus importantes sont dues aux données du corpus le plus court, on peut supposer qu'il n'y avait pas assez de données pour créer le modèles HMM des locuteurs et que les taux sont donc erronés pour la méthode utilisant la reconnaissance de parole. Ceci est confirmé par les coefficients de corrélation qui sont plus faibles pour la Table 3. 1 par rapport à la Table 3. 2.

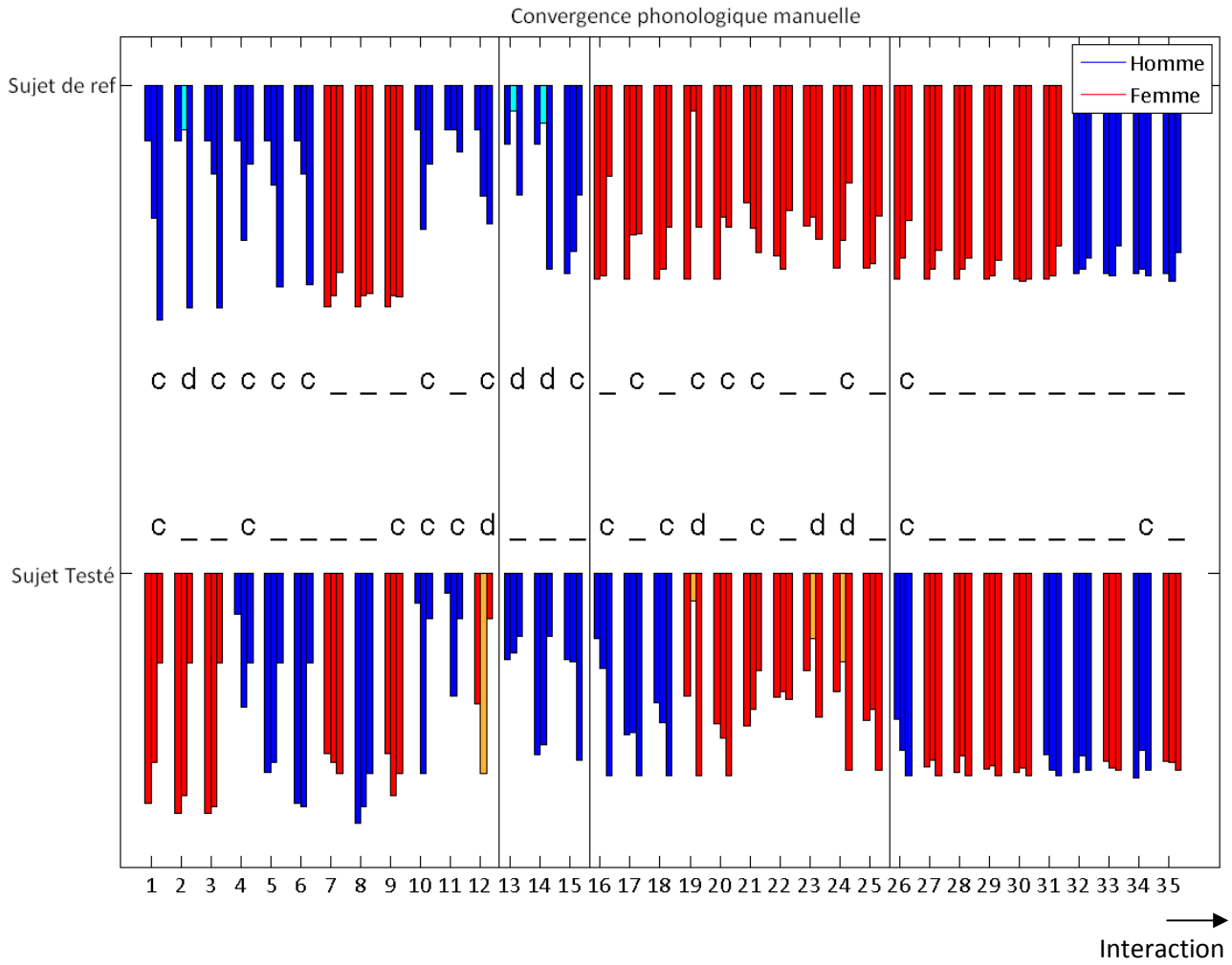


Figure 3. 3. Proportions de voyelles fermées prononcées par les sujets pendant leur pré-test (première barre), l'interaction (deuxième barre) et le pré-test de leur partenaire (troisième barre) pour 35 interactions. Ces proportions ont été calculées en utilisant l'alignement manuel. Les barres indiquées en bleu ciel ou en orange correspondent à des cas de divergence phonologique. Les étoiles correspondent à des convergences phonologiques supérieures à 10%.

Table 3. 1. Comparaison des deux méthodes utilisées pour l'étude de la convergence phonologique pour le corpus court (* = p<0.005, ** = p<0.01, * = p<0.05).**

Corpus court	Sujet de référence	Sujet testé
Propre pré-test	0.998 ***	0.706 *
Interaction	0.772 ***	0.625 *
Pré-test partenaire	0.712 **	0.903 ***

Table 3. 2. Comparaison des deux méthodes utilisées pour l'étude de la convergence phonologique pour le corpus long (* = p<0.005, ** = p<0.01, * = p<0.05).**

Corpus long	Sujet de référence	Sujet testé
Propre pré-test	0.999 ***	0.987 ***
Interaction	0.999 ***	0.998 ***
Pré-test partenaire	0.921 ***	0.968 ***

Pour la suite des analyses, nous avons d'abord utilisé la reconnaissance de parole pour aligner automatiquement les signaux avec le contenu linguistique, puis nous avons tout vérifié manuellement. Deux annotateurs ont vérifié les signaux pour minimiser les erreurs d'étiquetage. Nous devons cependant mentionner que l'étiquetage des variations allophoniques des voyelles moyennes est très difficile car les locuteurs français ont désormais tendance à avancer les voyelles moyennes fermées (Coveney, 2001; Boula de Mareüil *et al.,* 2008).

Chapitre 4 Caractérisation de la convergence phonétique

Comme nous l'avons dans le chapitre « Etat de l'art », l'accommodation mutuelle peut être étudiée à divers niveaux linguistiques : syntaxique, lexical, morphologique, phonologique et phonétique. Si l'échange de mots isolés permet de se concentrer sur les niveaux inférieurs en limitant les effets de contexte d'énonciation, il reste que les variantes de prononciation sont présentes. Bien que nous ayons sélectionné les dominos verbaux de manière à limiter les variantes accentuelles, nous avons observé quelques variations dialectales, particulièrement sur les variations allophoniques des voyelles moyennes.

Nous avons donc décidé a posteriori d'étudier le phénomène de convergence phonétique en mettant en correspondance les réalisations phonétiques de chaque allophone du système propre à chaque locuteur. Nous avons ainsi précisément étiqueté semi-automatiquement chaque mot de manière à pouvoir caractériser l'influence de l'interaction d'une part et le cas échéant sur le choix des variantes de prononciation et d'autre part sur les réalisations acoustiques de chaque allophone. Nous nommerons par la suite *convergence phonologique* le rapprochement de la distribution des choix de variantes de prononciation et *convergence phonétique* le rapprochement des caractéristiques acoustiques des réalisations de chaque catégorie phonétique.

4.1 Convergence phonétique

Nous avons dès lors considéré les variations des espaces sonores de chaque allophone de chaque interlocuteur entre son pré-test et son interaction avec un interlocuteur lors du jeu de dominos. Pour ceci nous rassemblons l'ensemble des réalisations portant la même étiquette phonétique.

Nous avons testé et comparé les résultats de la méthode de caractérisation la plus couramment utilisée dans la littérature – l'Analyse Discriminante Linéaire (LDA) utilisée par Delvaux et Soquet (2007) et par Aubanel (2011) – avec des méthodes issues de la reconnaissance de parole et du locuteur.

4.1.1 Analyse Discriminante Linéaire

La première méthode que nous avons choisie pour caractériser la convergence phonétique est l'analyse discriminante linéaire *(*Lelong et Bailly, 2011 ; Bailly et Lelong, 2010*).* Cette méthode répondait à nos objectifs car son but est de proposer un nouveau système de représentation tels que des individus du même groupe projetés sur ces axes soient le plus proches possibles les uns des autres, et que des individus de groupes différents soient le plus éloignés possible.

Delvaux et Soquet (Delvaux et Soquet, 2007) ont utilisé cette technique sur la durée des voyelles et les 20 premiers coefficients MFCC (Mel Frequency Cepstral Coefficients) des pré-tests de leurs paires de sujets. Aubanel et Nguyen (Aubanel et Nguyen, 2010) ont également utilisé l'analyse discriminante linéaire mais sur les coefficients DCT (Discrete Cosine Transform) des pré-tests. Les taux de

convergence trouvés restaient néanmoins faibles (de l'ordre de 10 % pour Delvaux et 20% pour Aubanel calculé dans le cas idéal) et très dépendants du contexte.

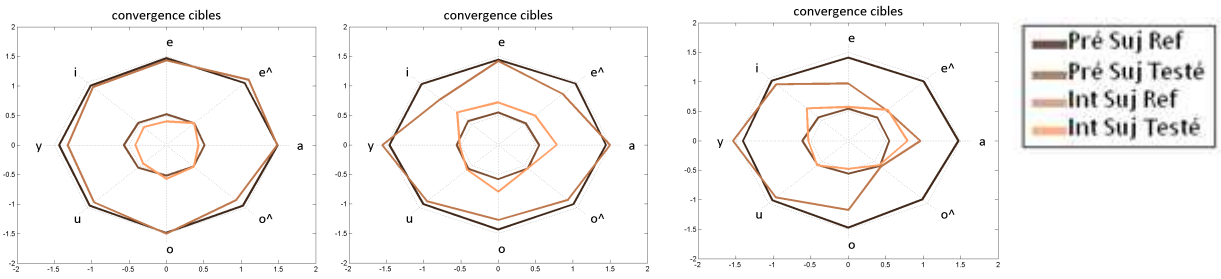
4.1.2 Méthode

Pré-traitement. Nous avons choisi d'utiliser l'analyse discriminante linéaire sur les coefficients MFCCs des voyelles [a], [ɛ], [e], [i], [y], [u], [o] et [ɔ]. Ces coefficients sont ceux qui décrivent le plus précisément le spectre de parole et qui se rapprochent le plus de la perception fréquentielle de l'oreille humaine. Nous avons extrait les 12 premiers coefficients MFCCs de chaque signal (étape 2 de la Figure 4. 2) en ne prenant pas en compte le premier coefficient qui correspond à l'énergie du signal. Ces coefficients ont été obtenus à l'aide de Praat (Boersma & Weenink, 2005*)* et calculé toutes les 10 ms en utilisant une fenêtre de 25 ms.

Correction canal. L'ensemble des MFCC des cibles vocaliques est normalisé par la moyenne et l'écart-type de chaque paramètre pour chaque corpus afin de s'affranchir des conditions d'enregistrement – microphone, dispositif d'enregistrement, température et acoustique de la pièce – bien que nous ayons pris soin de garder ces conditions expérimentales les plus constantes possibles.

Analyse discriminante linéaire. Nous avons sélectionné les coefficients correspondant aux voyelles en les prenant au milieu de l'alignement car on suppose que la voyelle est stable à cet endroit. Puis, nous avons utilisé les coefficients MFCC des premières moitiés de chaque pré-test comme paramètres d'entrées de l'analyse discriminante linéaire pour obtenir l'espace – i.e. premier axe discriminant – dans lequel les pré-tests des sujets de chaque paire étaient les plus éloignés (étape 4 de la Figure 4. 2).

Nous avons ensuite observé le taux de convergence moyens calculé sur les MFCCs pour chaque cible vocalique (voir Figure 4. 1). Sur la figure, les lignes marron correspondent aux taux de convergence calculés sur les moitiés des pré-tests de chaque sujet (condition de contrôle) alors que les lignes orangées correspondent au taux calculés en interaction. Ainsi, on aura une convergence si les lignes orangées se situent entre les deux lignes marron. On remarque que les résultats sont très dépendants des paires étudiées (voir (a) et (b)) ainsi que parfois des voyelles étudiées (voir (b) et (c)). Sur la figure (a) on observe une très faible adaptation des deux sujets alors que sur la figure (b), on peut voir qu'il y a une adaptation des deux sujets. Enfin la figure (c) montre que les taux de convergence sont très dépendants des voyelles étudiées, par exemple il n'y a pas de convergence pour le phonème [y] voire une divergence alors qu'il y a une adaptation mutuelle pour le phonème [ɛ]. Les taux trouvés restent cependant cohérents avec ce que nous attendions.



(a) (b) (c)

Figure 4. 1. Taux de convergence moyens pour chaque cible vocalique étudiée pour différentes paires. La ligne pointillée extérieure correspond au pré-test du sujet de référence et celle intérieure au pré-test du sujet testé. Les lignes pleines correspondent aux signaux d'interaction.

Après avoir validé chaque taux, nous avons calculé le taux de convergence moyen par sujet et par paire en prenant en compte l'ensemble des phonèmes étudiés.

Validation croisée. Après avoir utilisé HTK (Young *et al.*, 2009) pour obtenir une segmentation automatique de nos fichiers, nous les avons tous vérifiés manuellement. Nous avons ensuite séparé le pré-test de chaque sujet en deux parties égales de manière aléatoire (étape 1 de la Figure 4. 2). Nous avons répété cette opération 100 fois et calculé les taux de convergence moyens sur ces 100 itérations pour chaque paire afin d'obtenir une meilleure validation de nos résultats. Nous projetons ensuite les MFCCs des deuxièmes moitiés des pré-tests sur le premier axe discriminant afin d'obtenir la distance spectrale de référence pour chaque paire et ceux des signaux d'interaction pour observer comment cette distance a évolué (étape 4 de la Figure 4. 2). Comme nous ne projetons qu'une moitié du pré-test sur l'axe discriminant, nous comparons bien la même quantité de données entre le pré-test et l'interaction. La Figure 4. 2 résume la méthode utilisée.

Les taux de convergence ont été calculés à l'aide de la formule suivante :

$$C_{LDA\ 1\rightarrow 2} = \frac{P_{l1} - I_{l1l2}}{P_{l1} - P_{l2}} \qquad (4.1)$$

Où I_{l1l2} correspond au signal d'interaction du sujet 1 avec le sujet 2, P_{l1} et P_{l2} correspondent respectivement aux pré-tests du sujet 1 et du sujet 2. Ainsi si on n'observe pas de convergence phonétique, I_{l1l2} sera égal à P_{l1} et le taux de convergence vaudra alors 0. Alors que s'il y a une adaptation complète, I_{l1l2} sera égal à P_{l2} et le taux de convergence sera égal à 1. La formule est donc cohérente.

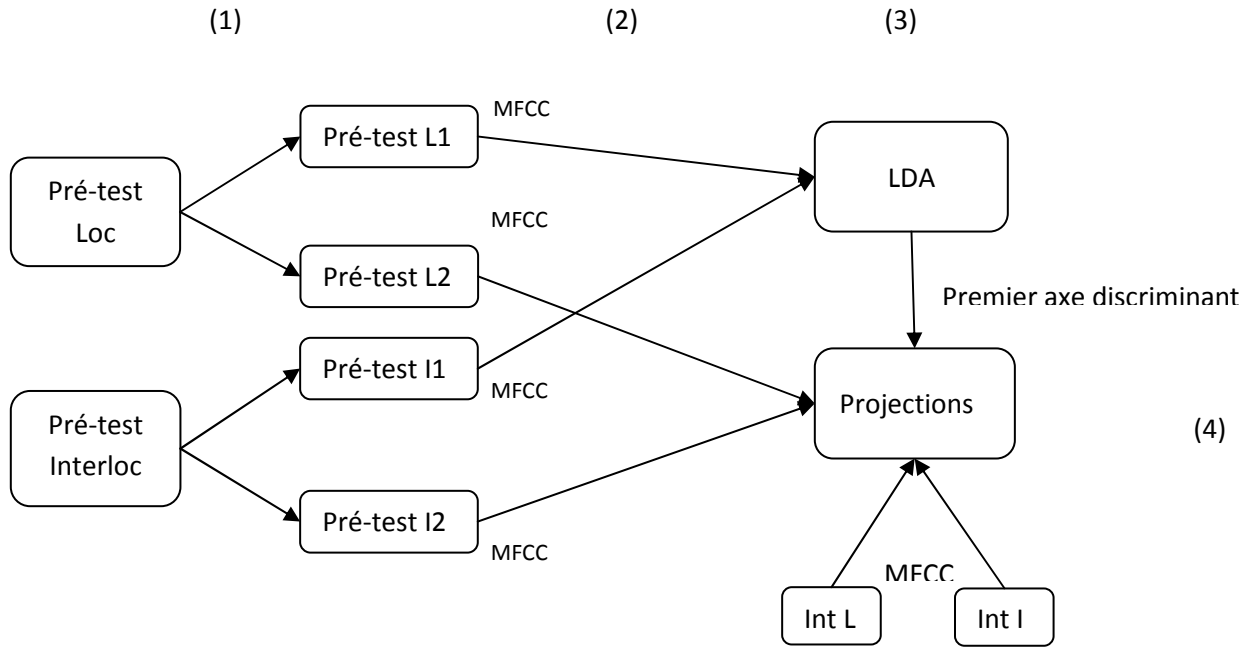


Figure 4. 2. Description de la méthode utilisée avec l'analyse discriminante linéaire sur les coefficients MFCC, L correspond à locuteur ou sujet de référence, I correspond à interlocuteur ou sujet testé.

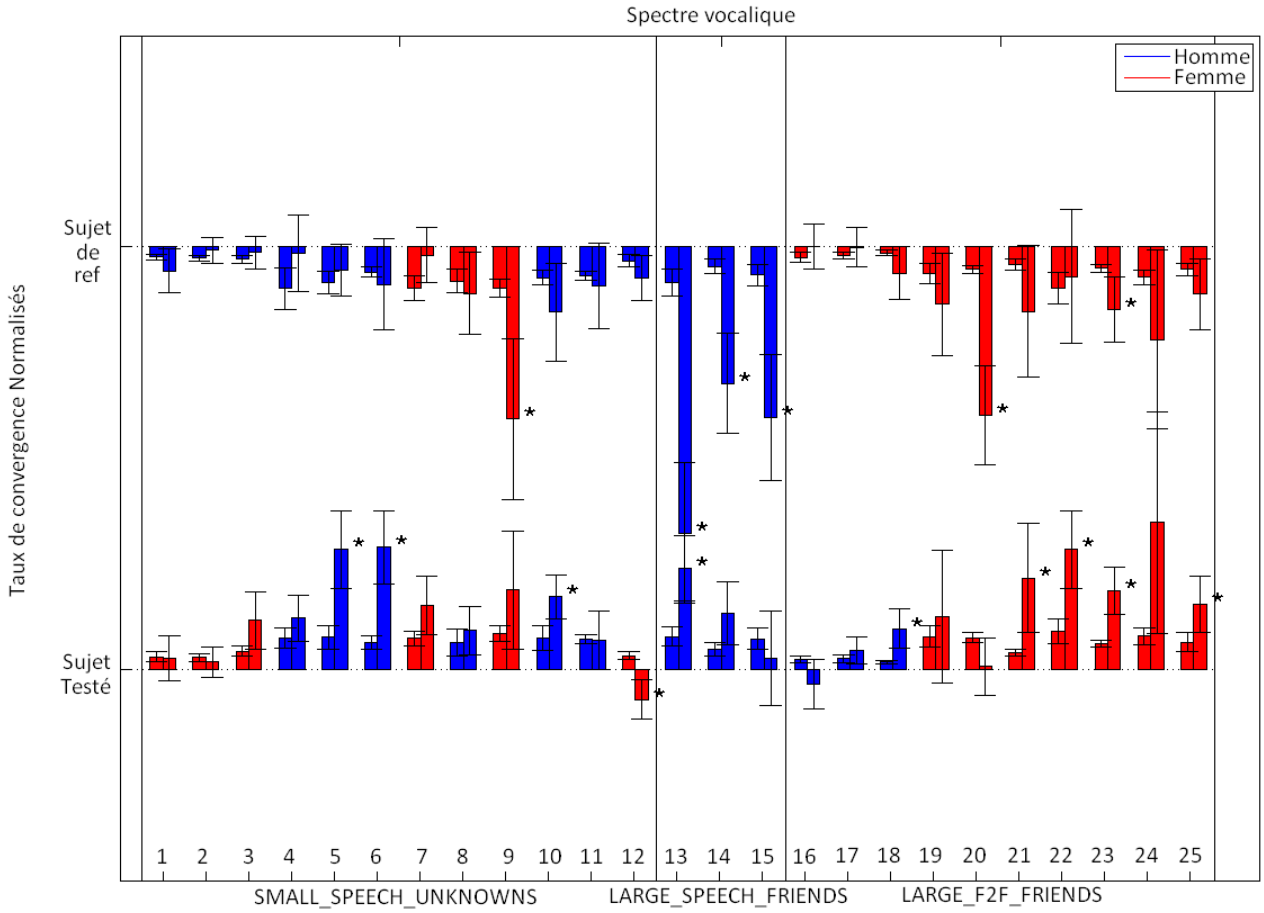


Figure 4. 3. Taux de convergence moyen calculé sur 100 itérations des cibles vocaliques des interlocuteurs pour les expériences I, II et III. Une analyse discriminante linéaire a été utilisée sur une moitié, aléatoirement décidée, des pré-tests pour séparer au maximum les espaces vocaliques des sujets. Nous obtenons ainsi une distance de référence entre les sujets qui est utilisée pour calculer des taux de convergence normalisés. Nous l'utilisons d'abord pour calculer le taux de convergence sur l'autre moitié du pré-test pour obtenir un point de référence pour chaque interlocuteur (colonne de gauche). Les taux de convergence des signaux d'interactions sont représentés sur la colonne de droite. On a inversé les taux de convergences du sujet de référence pour mettre en relief le rapprochement des sujets. Les distributions marquées par une étoile sont significativement différentes (p<0.1) du pré-test correspondant.

4.1.3 Résultats

4.1.3.1 Convergence globale

La Figure 4. 3 montre les résultats observés pour les expériences I, II et III. On aura une adaptation des sujets si le sujet testé – représenté en ordonnée 0 – se rapproche du sujet de référence – représenté en ordonnée 1 – et/ou réciproquement. On remarque bien une adaptation des sujets mais celle-ci n'est pas systématique.

Expérience I. On peut voir sur l'expérience I (paires 1 à 12) que le phénomène est amplifié avec des paires de même sexe et particulièrement entre femmes (cf paire 9), cela peut être dû à une similarité plus importante entre les voix des femmes, comme souligné par Pardo (2006) et bien d'autres auteurs. Nous observons également que la convergence obtenue n'est pas très importante. On rappelle que

pour l'expérience I, les sujets ne se connaissaient pas. On suppose alors que l'amplitude du phénomène reste faible car les sujets ne possèdent pas encore de modèles internes de leur interlocuteur. Cela peut également s'expliquer par le fait qu'il n'y ait pas de contact visuel entre les sujets ce qui implique un engagement moindre dans l'interaction. Nous avons donc décidé de faire participer des paires d'amis pour les expériences suivantes.

Expérience II. Pour l'expérience II (paire 13 à 15), on remarque bien une convergence plus importante pour des paires de même sexe qui se connaissent depuis plusieurs années. Il s'agit ici de trois paires d'hommes. Mais dans ce cas les sujets interagissaient toujours en face-à-face médiatisé et sans contact visuel.

Expérience III. Pendant l'expérience III (paire 16 à 25), les sujets interagissent en face-à-face et se connaissent. Nous avons fait une première série de test avec des paires mixtes (paires 16 à 18) pour confirmer notre hypothèse sur les paires de même sexe, puis nous avons exclusivement enregistré des paires d'amies (paires 19 à 25) pour lesquelles nous avons bien observé une convergence plus systématique.

Nous avons testé le facteur Conditions – Pré-test vs. Interaction – avec une ANOVA pour voir si les distributions des scores de convergence étaient significativement différentes entre le pré-test – à ce moment, ils doivent être proche de 0 – et l'interaction. Celles qui l'étaient ($p<0.1$) ont été dessinées en gras. On remarque que 5 interactions (paires 20, 21, 22, 23 et 25) sur 7 entre des femmes qui se connaissent depuis quelque temps montrent une convergence significative.

La Table 4. 1 résume les taux de convergence moyens par paire pour le pré-test et l'interaction, ainsi que l'écart type des distributions et le résultat de l'ANOVA. Les distributions de l'interaction significativement différentes de celles des pré-tests sont indiquées en gras. Les taux de convergence obtenus sont compris entre -0,04 et 0,41 pour l'expérience I entre des inconnus et entre -0,03 et 0,41 pour les expériences II et III entre des amis. Cependant, on remarque bien un phénomène de convergence plus systématique pour les expériences dont les participants se connaissaient et ceci est encore plus vrai pour les paires de même sexe.

La Figure 4. 4 illustre 4 spectrogrammes démontrant la convergence. Le premier spectrogramme (en haut à gauche) représente le mot « gerçure » prononcé par le locuteur 1 pendant son pré-test alors que la quatrième (en bas à droite) représente ce même mot prononcé par le locuteur 2 pendant son pré-test. Le deuxième spectrogramme (en haut à droite) représente le mot « gerçure » prononcé par le locuteur 1 en interaction avec le locuteur 2 et le troisième spectrogramme représente le même mot prononcé par le locuteur 2 en interaction avec le locuteur 1. On observe un alignement au niveau de la durée des mots prononcés en interaction par rapport à ceux prononcés pendant les pré-tests. On peut également voir que le ε du locuteur 1 prononcé en interaction est plus proche de celui du locuteur 2 en interaction ou en pré-test.

Table 4. 1. Tableau récapitulatif des taux de convergence des expériences I, II et III

			Sujet de référence					Sujet testé					
Exp	Paires	Sexe	m Pré	std Pré	m Inter	std Inter	p	Sexe	m Pré	std Pré	m Inter	std Inter	p
	1	H	0,02	0,02	0,06	0,1	0,4	F	0,03	0,03	0,03	0,11	0,95
	2	H	0,03	0,01	0	0,06	0,44	F	0,03	0,02	0,02	0,07	0,69
I	3	H	0,03	0,02	0,01	0,08	0,59	F	0,04	0,02	0,12	0,14	0,15
186	4	H	0,1	0,1	0,02	0,18	0,27	H	0,07	0,05	0,12	0,11	0,29
Dominos	5	H	0,08	0,05	0,06	0,12	0,56	H	**0,08**	**0,06**	**0,28**	**0,19**	**0,01**
médiatisé	6	H	0,06	0,02	0,08	0,22	0,71	H	**0,06**	**0,03**	**0,3**	**0,17**	**0**
Inconnus	7	F	0,1	0,06	0,02	0,13	0,14	F	0,07	0,03	0,15	0,14	0,15
	8	F	0,08	0,06	0,12	0,19	0,69	H	0,06	0,06	0,09	0,11	0,56
	9	F	**0,1**	**0,04**	**0,41**	**0,38**	**0,04**	F	0,08	0,04	0,19	0,28	0,32
	10	H	0,07	0,03	0,15	0,23	0,33	H	**0,07**	**0,06**	**0,17**	**0,1**	**0,04**
	11	H	0,07	0,02	0,09	0,2	0,74	H	0,07	0,02	0,07	0,14	0,97
	12	H	0,03	0,03	0,07	0,11	0,32	F	**0,03**	**0,02**	**-0,07**	**0,09**	**0,01**
II	13	H	**0,09**	**0,06**	**0,68**	**0,33**	**0**	H	**0,08**	**0,05**	**0,24**	**0,15**	**0,01**
médiatisé	14	H	**0,05**	**0,03**	**0,32**	**0,24**	**0,006**	H	0,05	0,03	0,13	0,15	0,13
Amis	15	H	**0,07**	**0,05**	**0,4**	**0,3**	**0,01**	H	0,07	0,05	0,03	0,22	0,59
	16	F	0,02	0,02	0	0,11	0,51	H	0,02	0,02	-0,03	0,12	0,19
	17	F	0,02	0,02	0	0,09	0,57	H	0,03	0,02	0,05	0,07	0,44
	18	F	0,01	0,01	0,06	0,12	0,3	H	**0,02**	**0,01**	**0,1**	**0,09**	**0,03**
III	19	F	0,06	0,05	0,14	0,24	0,42	F	0,08	0,05	0,12	0,31	0,68
350	20	F	**0,05**	**0,02**	**0,4**	**0,23**	**0**	F	0,07	0,02	0,01	0,14	0,19
Dominos	21	F	0,04	0,03	0,15	0,31	0,33	F	**0,04**	**0,02**	**0,22**	**0,26**	**0,07**
Amis	22	F	0,1	0,07	0,07	0,32	0,81	F	**0,09**	**0,06**	**0,28**	**0,18**	**0,01**
	23	F	**0,05**	**0,02**	**0,15**	**0,15**	**0,09**	F	**0,06**	**0,02**	**0,18**	**0,11**	**0,01**
	24	F	0,07	0,03	0,22	0,42	0,34	F	0,08	0,04	0,35	0,52	0,17
	25	F	0,05	0,03	0,11	0,17	0,34	F	**0,06**	**0,04**	**0,15**	**0,13**	**0,09**

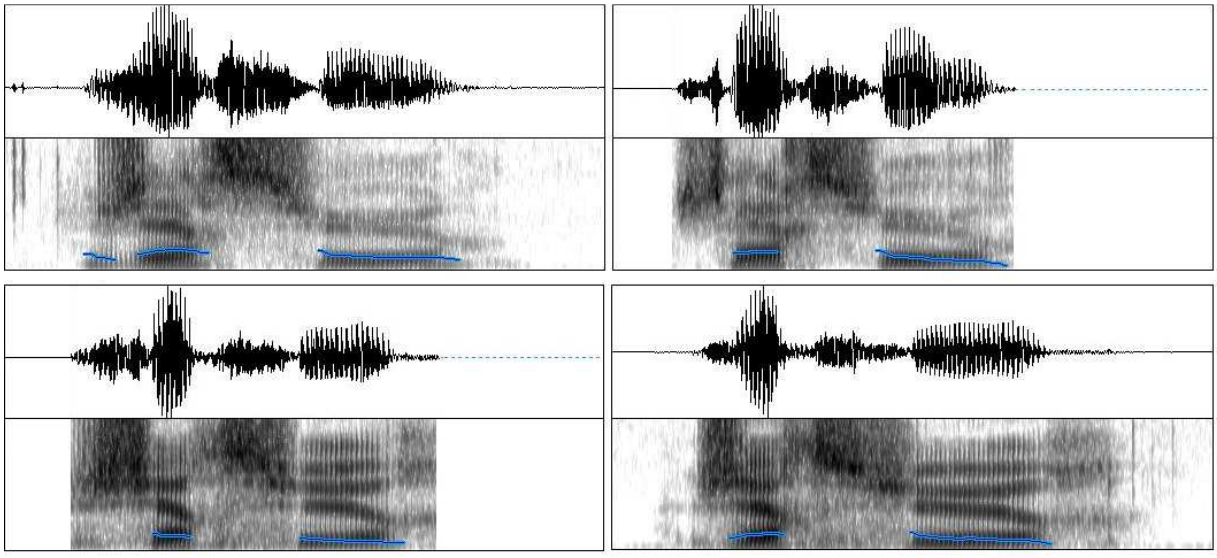


Figure 4. 4. Spectrogrammes du mot « gerçure » prononcé par le locuteur 1 pendant son pré-test (en haut à gauche), par le locuteur 1 en interaction avec le locuteur 2 (en haut à droite), par le locuteur 2 en interaction avec le locuteur 1 (en bas à gauche) et par le locuteur 2 pendant son pré-test (en bas à droite). On observe que les durées des items en interaction sont plus courtes que celles de ceux en pré-test. On remarque également des similitudes au niveau du ɛ pendant l'interaction.

4.1.3.2 Convergence vocalique

Comme nos sujets de référence ont interagi avec différents interlocuteurs, nous avons également représenté les ellipses de dispersions dans le premier plan discriminant pour chaque voyelle (voir Figure 4. 5). On remarque une nouvelle fois que les résultats sont très dépendants des paires étudiées. Sur la Figure 4. 5, on observe alb en interaction avec trois interlocuteurs différents A, B et C. L'ellipse sombre au centre ainsi que les trois ellipses pleines en périphérie correspondent réciproquement aux pré-tests d'alb et de ces trois interlocuteurs. Les ellipses vides au centre de la figure représentent les MFCCs d'alb en interaction avec chacun des sujets A, B et C alors que les autres ellipses pleines correspondent aux signaux d'interaction des trois sujets testés. On peut voir que, suivant les interactions, on obtient des résultats différents. Alors qu'A et B s'adaptent complètement à alb, C ne bouge pas par rapport à son pré-test. Cela peut s'expliquer par le fait qu'alb, A et B sont des femmes et C est un homme ce qui confirme que la convergence est plus facile entre paire de même sexe et plus particulièrement pour les femmes (Namy *et al.*, 2002).

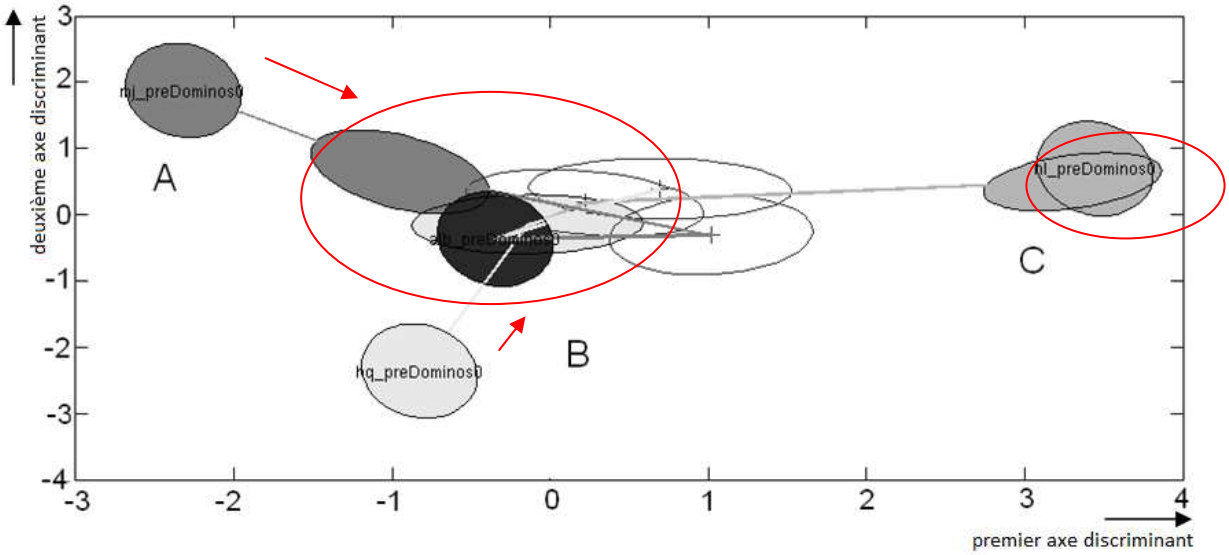


Figure 4. 5. Projection sur le premier espace discriminant des MFCCs de la cible vocalique [ɔ] produit par le locuteur (ellipse de dispersion sombre pour le pré-test au centre de la figure) en interaction avec trois interlocuteurs A, B, C (les ellipses de leur pré-test se situent à la périphérie). Les réalisations pour les interactions sont indiquées avec des ellipses vides pour alb et des ellipses remplies avec la même couleur que le pré-test pour les interlocuteurs. On remarque qu'A et B converge vers alb alors que C ne s'adapte pas du tout.

4.1.4 Commentaires

Ces résultats démontrent que les sujets adaptent leur contenu phonétique en fonction de leur interlocuteur mais que cette adaptation dépend de nombreux facteurs comme le lien social entre les sujets (inconnus vs. amis), le sexe des sujets (paires mixtes ou non, Hommes vs. Femmes) ou le rôle des sujets (relation de dominance/connaissance de la tâche dans la conversation). Cela nous a amené à tester notre paradigme sur des sujets provenant d'une même famille pour lesquelles l'expérience commune est plus importante. On s'attend donc à ce que les taux de convergence soient plus élevés dans cette étude. Elle sera expliquée à la partie (voir §4.5).

Nous avons également voulu trouver une méthode rapide et automatique pour caractériser la convergence, nous avons donc utilisé la reconnaissance du locuteur.

La méthode utilisée se focalise uniquement sur un taux de convergence obtenu à partir des voyelles. On ne prend donc pas en compte le phénomène de coarticulation. De plus, comme nous utilisons des mots disyllabiques, nous comparons des voyelles qui ne sont pas forcément à la même place dans le mot – i.e. syllabe initiale ou finale –, elles vont donc être accentuées de manière différente. Nous avons donc développé une autre méthode nous permettant de nous affranchir de toutes ces contraintes et d'obtenir ainsi un taux de convergence globale. Nous avons opté pour la reconnaissance du locuteur (voir §4.2).

4.2 Reconnaissance du locuteur

La reconnaissance du locuteur consiste à distinguer différents locuteur en fonction de leur signature vocale. On distingue en général l'identification et la vérification du locuteur. L'identification consiste à reconnaître un locuteur appartenant à une population composée de plusieurs locuteurs en comparant son expression vocale à des références connues. La vérification consiste à accepter ou refuser une identité attribuée à un locuteur. Pour cela, on étudie la distance entre ses caractéristiques vocales pendant le test et pendant son enregistrement de référence et on compare cette distance à un seuil d'acceptation.

Dans notre cas, nous utilisons les modèles de mélange de gaussiennes car cette méthode nous permet d'obtenir un modèle probabiliste de nos locuteurs. Nous obtiendrons de meilleures performances avec ces modèles car ils prendront en compte la variabilité intra-locuteur, – i.e un même locuteur ne prononce pas deux fois la même phrase de la même manière – et la variabilité du canal dû à différentes conditions d'enregistrements.

4.2.1 Les modèles de mélange de gaussiennes (GMM)

Les modèles de mélange de gaussiennes ou GMM pour *Gaussian Mixture Model* sont utilisés en reconnaissance du locuteur car ils permettent de modéliser un locuteur par une somme pondérée de composantes gaussiennes (Reynolds, 1995). Une combinaison linéaire de gaussiennes permet alors de représenter une large variété de distributions.

Avec les modèles de mélange de gaussiennes, on considère que chaque observation y est la réalisation d'une variable aléatoire dont la densité de probabilité est notée $p(y|\theta)$, θ correspond à l'ensemble des paramètres du modèles. Cette densité de probabilité est décrite comme une somme de M composantes (M étant le nombre de gaussiennes définie empiriquement) de la manière suivante :

$$p(y|\theta) = \sum_{i=1}^{M} \beta_i p_i(y|\theta_i) \qquad (4.2)$$

Avec $y = [y_1, y_2, ..., y_d]$, d étant la dimension du vecteur de caractéristiques, $\theta_i = \{\mu_i, \Sigma_i\}$ étant les paramètres d'une distribution gaussienne de moyenne $\mu_i = [\mu_{i1}, \mu_{i2}, ..., \mu_{id}]$ et de matrice de covariance Σ_i.

Comme $p_i(y/\theta)$ est une distribution gaussienne, elle peut s'écrire sous la forme :

$$N(y, \mu_i, \Sigma_i) = p_i(y|\theta_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} e^{-\frac{1}{2}(y-\mu_i)^T \Sigma_i^{-1}(y-\mu_i)} \quad (4.3)$$

De plus, les poids β_i, définis pour chaque composantes gaussiennes, vérifient les lois suivantes :

$$\sum_{i=1}^{M} \beta_i = 1 \quad et \quad \forall\, i \in [1; M], \beta_i \geq 0 \quad (4.4)$$

Le modèle de mélange de gaussiennes sera alors complètement déterminé par les M vecteurs de moyennes, les M matrices de covariance et les M poids β_i.

La méthode traditionnelle pour déterminer ces paramètres est l'algorithme « EM » (*Expectation-Maximization* en Anglais) introduite par Dempster (Dempster *et al*, 1977). Pour déterminer les paramètres qui définissent le GMM de chaque locuteur, on cherche à maximiser le logarithme de la fonction de vraisemblance du modèle notée $L(\theta)$ et définie par :

$$L(\theta) = \sum_{t=1}^{N_T} log\, p(y_t|\theta) \quad (4.5)$$

Où $Y = [y_1, y_2, \ldots, y_{NT}]$ correspond à une séquence de N_T vecteurs d'entraînements.

Le problème est que cette expression n'est pas linéaire en fonction des paramètres $\theta = \{\mu, \Sigma, \beta\}$ qui vont définir le GMM. On utilise donc l'algorithme « EM » pour résoudre itérativement ce problème. Chaque itération implique deux étapes successives, une étape d'estimation et une étape de maximisation. L'idée est qu'en partant d'un modèle initial (déterminé de manière aléatoire), on va estimer un nouveau modèle tel que la log vraisemblance $L(\theta)$ augmente. Ce nouveau modèle deviendra ensuite le modèle initial pour la prochaine itération et le processus se répète jusqu'à ce qu'un seuil de convergence soit atteint. L'étape d'estimation va permettre de calculer, à partir d'un modèle $\theta^{(m)}$ obtenu à la m-ième itération, la probabilité conditionnelle a posteriori qu'une observation y_t provienne de la i-ème gausienne notée g_i. Cette probabilité est donnée par la fonction suivante :

$$P(g_i|y_t, \theta^m) = \frac{\beta_i p_i\left(y_t \middle| \theta_i^m\right)}{\sum_{j=1}^{M} \beta_j p_j\left(y_t \middle| \theta_j^m\right)} \quad (4.6)$$

On peut alors déterminer les paramètres du prochain modèle $\theta^{(m+1)}$ à partir de cette quantité grâce aux formules suivantes :

$$\beta_i^{m+1} = \frac{1}{N_T} \sum_{t=1}^{N_T} P(g_i|y_t, \theta^m) \quad (4.7)$$

$$\mu_i^{m+1} = \frac{\sum_{t=1}^{N_T} P(g_i|y_t, \theta^m)\, y_t}{\sum_{t=1}^{N_T} P(g_i|y_t, \theta^m)}$$

$$\Sigma_i^{m+1} = \frac{\sum_{t=1}^{N_T} P(g_i|y_t, \theta^m)\, (y_t - \mu_i^m)(y_t - \mu_i^m)^T}{\sum_{t=1}^{N_T} P(g_i|y_t, \theta^m)}$$

Après plusieurs itérations, on trouve les paramètres $\theta = \{\mu, \Sigma, \beta\}$ qui définiront chaque GMM pour chaque locuteur.

4.2.2 Méthode

Nous avons utilisé cette méthode pour caractériser la convergence car elle nous permet d'utiliser une analyse de données qui est indépendante du texte prononcé. Cela permettra alors de mettre en place des paradigmes en conditions moins contrôlées pour étudier le phénomène de convergence phonétique. Les modèles de chaque locuteur ont été créés grâce à la plateforme Alizée (Charton *et al.,* 2010), développée au Laboratoire d'Informatique d'Avignon.

La Figure 4. 6 décrit l'algorithme que nous avons utilisé pour obtenir le modèle GMM de chaque interlocuteur. Nous avons d'abord séparé le pré-test de chaque locuteur en deux parties égales (voir étape (1) sur la Figure 4. 6). Puis nous avons supprimé les parties de silence pour éviter d'obtenir une modélisation de l'environnement. La première partie nous a permis de créer le modèle GMM de chaque locuteur alors que la deuxième partie a été utilisée comme test de manière à obtenir une distance inter-locuteur de référence. Nous avons ensuite extrait les paramètres cepstraux (voir étape (2) de la Figure 4. 6) de chaque signal à l'aide du logiciel Spro (calculés toutes les 10 ms à partir d'une fenêtre de 20 ms).

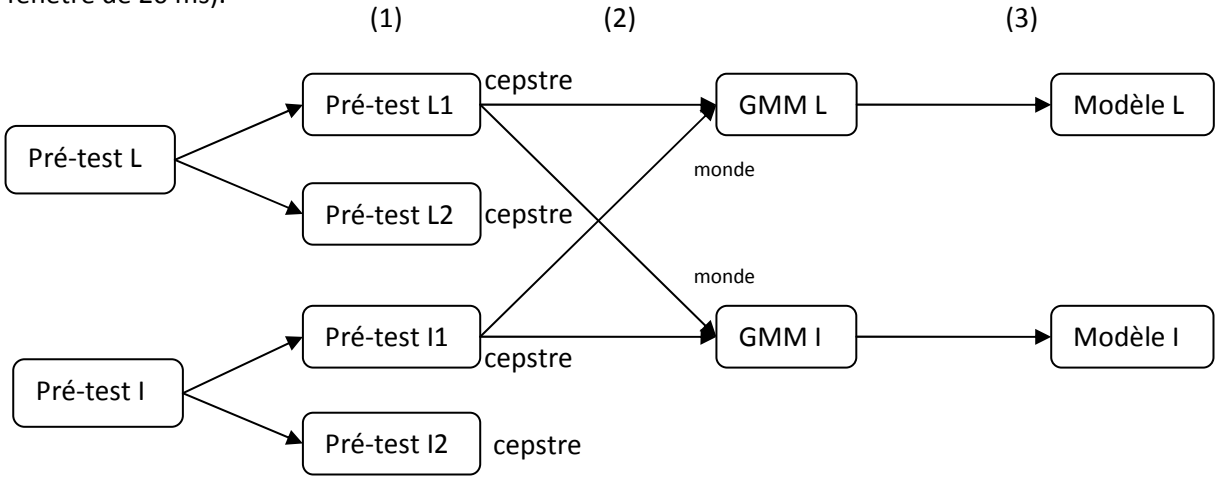


Figure 4. 6. Algorithme utilisé pour créer un modèle GMM de chaque locuteur à partir de leur pré-test. Pendant l'étape (1), nous séparons les pré-tests en deux parties égales, puis dans l'étape (2) nous extrayons les paramètres cepstraux grâce à Spro, enfin, dans l'étape (3), nous entraînons le modèle GMM de chaque locuteur en prenant en compte les paramètres d'une première partie du pré-test du locuteur (la deuxième servira pour le test) et les paramètres d'une partie du pré-test de l'interlocuteur qui servira à construire le « monde » (les voix différentes de celles du locuteur testé).

Pour chaque GMM, nous avons d'abord entraîné le modèle du « monde », celui représentant le modèle universel de locuteurs. Pour chaque locuteur, nous avons défini le « monde » à partir du signal de pré-test de son interlocuteur (ce signal a également été divisé en deux parties égales ce qui nous permettra d'utiliser la validation croisée). Nous avons choisi de créer des GMMs à 64 composantes. Ce paramètre ainsi que tous les autres ont été déterminés empiriquement de manière à maximiser la distance de référence entre les espaces acoustiques des locuteurs de chaque paire ($LLR_{l1l2}(P_{l1})+LLR_{l2l1}(P_{l2})$ sur l'ensemble des trames P_{l1} et P_{l2} prononcées respectivement par les locuteurs l1 et l2 pendant le pré-test). Nous avons commencé par prendre la moitié des trames du signal pour initialiser notre modèle de GMM à l'aide d'une quantification vectorielle puis nous avons utilisé 10 itérations de l'algorithme « EM » pour obtenir un modèle du « monde ». Nous avons ensuite entraîné le modèle chaque locuteur en utilisant 5 itérations en adaptant le modèle du monde via la méthode du maximum à posteriori (MAP) (voir l'étape (3) de la Figure 4. 6).

Après avoir obtenu un modèle de chaque locuteur de chaque paire, nous avons calculé la log-vraisemblance (*LLR pour Log Likelihood Ratio*) de la partie restante du pré-test de chaque interlocuteur avec les deux modèles obtenus pour l'interaction en question. Nous avons procédé au même calcul avec les signaux produits pendant l'interaction (voir Figure 4. 7). Pour calculer la vraisemblance d'un signal y, nous utilisons deux hypothèses :

H_S : Le signal y a été produit par le locuteur S

H_{-S} : Le signal y a été produit par le partenaire du locuteur S

Le rapport de vraisemblance est alors calculé de la manière suivante :

$$LR_s(Y) = \prod_{y \in Y} \frac{p(y|H_s)}{p(y|H_{-s})} < \delta \tag{4.8}$$

Où $p(y|H)$ est la fonction de densité de probabilité pour l'hypothèse H, évaluée pour le segment de parole y et δ est le seuil de décision pour accepter ou rejeter l'hypothèse H_S.

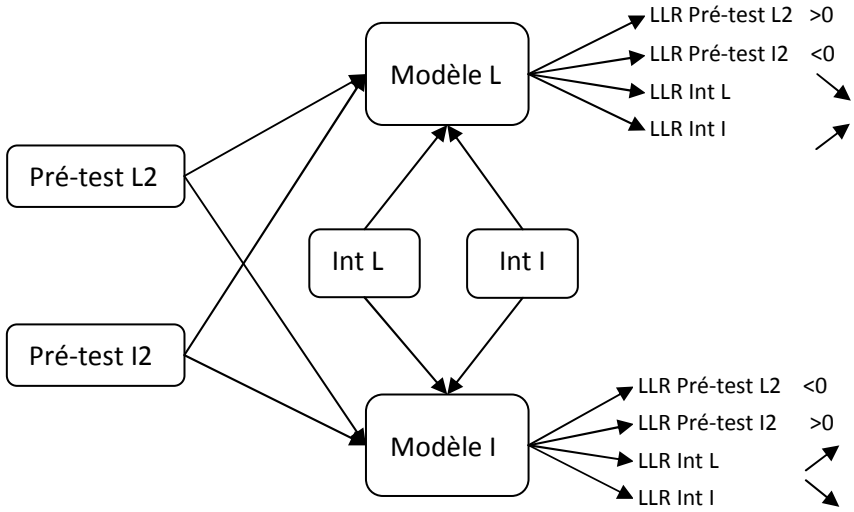


Figure 4. 7. Calculs de la log-vraisemblance des signaux de pré-test et d'interaction en utilisant les modèles de chaque interlocuteur.

Avec Alizée, on calcule le rapport de log-vraisemblance à partir d'un ensemble de trames test Y. Deux GMMs décrivent respectivement $p(y|H_s)$ et $p(y|H_{\neg s})$. Dans notre cas, H_S et $H_{\neg S}$ sont les modèles des deux locuteurs de chaque paire. On calcule alors la log-vraisemblance de la façon suivante :

$$LLR_{l1l2}(Y) = \sum_{y \in Y} log\left(\frac{p(y|H_{l1})}{p(y|H_{l2})}\right) \quad (4.9)$$

Le taux de convergence de l1 vers l2, noté $C_{LLR1\rightarrow 2}$ est alors calculé comme le quotient relatif entre la différence du LLR d'un locuteur (ici *l1*) calculé avec son propre modèle sur les trames du pré-test P_{l1} et pendant l'interaction (I_{l1l2}) et la différence de LLR calculé sur les trames des pré-tests (P_{l1} et P_{l2}).

$$C_{LLR1\rightarrow 2} = \frac{LLR_{l1l2}(P_{l1}) - LLR_{l1l2}(I_{l1l2})}{LLR_{l1l2}(P_{l1}) - LLR_{l1l2}(P_{l2})} \quad (4.10)$$

Où I_{l1l2} est l'ensemble des trames prononcées par le locuteur l1 pendant son interaction avec l2. Ainsi, s'il n'y a pas de convergence observée, on obtient alors $I_{l1l2}=P_{l1}$ et $C_{LLR1\rightarrow 2}= 0$. Notre définition est donc cohérente. Comme le pré-test a été divisé en deux parties, on compare la même quantité de données entre le pré-test et l'interaction. Le tableau suivant résume les résultats attendus en cas de convergence.

	Pré-test L	Pré-test I	Interaction L	Interaction I
Modèle L	LLR >0	LLR <0	LLR >0, ↘	LLR <0, ↗
Modèle I	LLR <0	LLR >0	LLR <0, ↗	LLR >0, ↘

Table 4. 2. Tendance des résultats en cas d'adaptation des sujets. La vraisemblance d'un locuteur calculée avec son propre modèle est positive alors qu'elle est négative lorsqu'on la calcule avec le modèle d'un autre locuteur. Ainsi, on a une convergence d'un sujet si la vraisemblance calculée sur le signal d'interaction de celui-ci diminue par rapport à celle de son pré-test lorsqu'on utilise son propre modèle et augmente lorsqu'on utilise le modèle de son interlocuteur (i.e. ses caractéristiques vocales sont plus semblables à celles de son interlocuteur pendant l'interaction). La convergence se traduit alors par le couplage de deux phénomènes : l'éloignement de ses propres références et le rapprochement des réalisations à l'espace de référence de l'autre.

La Figure 4. 8 représente les rapports de log-vraisemblance croisés pour chaque interaction. Les grandes croix à gauche représentent le LLR calculé sur le pré-test d'un locuteur avec son propre modèle alors que les petites croix représentent le LLR calculé avec le modèle de son interlocuteur. Les symboles décalés sur la droite représentent la même chose mais calculés pendant l'interaction. La distance entre chaque grand et petit symbole traduit la distance entre les espaces acoustiques de chaque locuteur réciproquement pendant le pré-test et pendant l'interaction. On observe une symétrie dans les comportements de chaque locuteur de chaque paire due à la manière dont le « monde » a été crée (i.e. à partir du pré-test de l'interlocuteur).

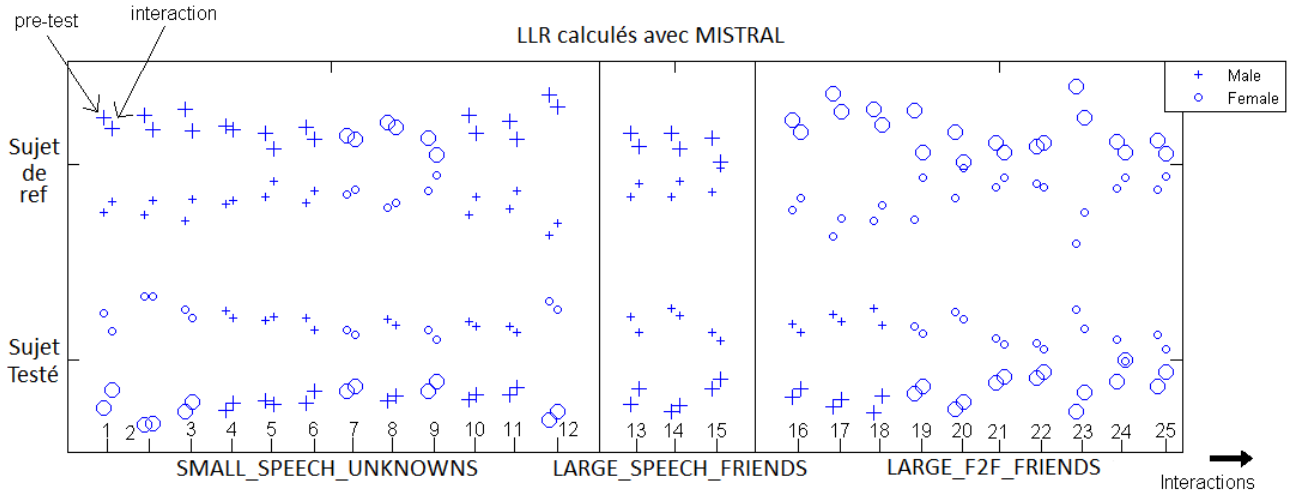


Figure 4. 8. LLR calculés pour chaque locuteur avec son propre modèle (en grand) et avec le modèle de son interlocuteur (en petit) pendant le pré-test de chacun (à gauche) et pendant l'interaction (à droite). Les croix représentent les hommes et les cercles représentent les femmes. La distance entre les grands et les petits symboles correspond à la distance inter-locuteurs. Les scores du locuteur de référence ont été inversés pour mettre en relief la symétrie due à la méthode utilisée pour construire le « monde ».

Pour obtenir une meilleure validation de nos résultats, nous avons utilisé une validation croisée sur quatre conditions et calculé des taux de convergence en faisant la moyenne des taux de convergence obtenus pour chaque condition. La Table 4. 3 résume les quatre conditions utilisées.

	Modèle	Monde	Test	Test
Expérience 1	Pré-test 1 du locuteur	Pré-test 1 de l'interlocuteur	Pré-test 2 du locuteur	Interaction du locuteur
Expérience 2	Pré-test 1 du locuteur	Pré-test 2 du locuteur	Pré-test 2 du locuteur	Interaction du locuteur
Expérience 3	Pré-test 2 du locuteur	Pré-test 1 de l'interlocuteur	Pré-test 1 du locuteur	Interaction du locuteur
Expérience 4	Pré-test 2 du locuteur	Pré-test 2 du locuteur	Pré-test 1 du locuteur	Interaction du locuteur

Table 4. 3. Présentation des quatre expériences définies pour pouvoir utiliser une validation croisée de nos résultats. Nous calculons d'abord la vraisemblance sur le pré-test (3ème colonne) ce qui nous permet d'obtenir le « point de départ » pour chaque locuteur et donc de normaliser notre taux de convergence grâce à la vraisemblance calculée sur le signal d'interaction (4ème colonne).

4.2.3 Résultats

La méthode de caractérisation de la convergence phonétique à partir de la reconnaissance du locuteur permet d'obtenir une mesure objective de la convergence sans utiliser la segmentation du corpus.

La Table 4. 4 regroupe les résultats obtenus avec cette méthode sur les trois types d'expériences étudiés. On remarque que les taux de convergence sont plus élevés lorsque les sujets testés sont amis (0.21 vs 0.13). Les scores de convergence restent faibles pour la condition impliquant des Inconnus, on obtient une faible divergence pour le sujet testé de la paire 5 puis les taux de convergence varient entre 0.04 (paire 4) et 0.3 (paire 9). Pour les deux autres conditions, on observe également une divergence pour le sujet de référence de la paire 22 mais les taux de convergence sont compris cette fois entre 0.08 (paire 20) et 0.54 (paire 24).

			Sujet de référence			Sujet Testé	
Exp	Paires	Sexe	m Pre	m Inter	Sexe	m Pre	m Inter
	1	H	0	0,11	F	-0,01	0,18
	2	H	0	0,15	F	-0,01	-0,01
I	3	H	0	0,2	F	-0,01	0,08
186	4	H	0	0,04	H	0	0,07
Dominos	5	H	0	0,26	H	-0,01	-0,07
médiatisé	6	H	0	0,16	H	-0,01	0,15
Inconnus	7	F	-0,02	0,1	F	-0,02	0,07
	8	F	-0,02	0,06	H	-0,01	0,07
	9	F	-0,02	0,25	F	-0,02	0,15
	10	H	0	0,17	H	-0,01	0,07
	11	H	0	0,2	H	-0,02	0,09
	12	H	0	0,09	F	0	0,06
II	13	H	-0,01	0,21	H	0	0,17
médiatisé	14	H	-0,01	0,25	H	0	0,06
Amis	15	H	0	0,4	H	-0,01	0,14
	16	F	0	0,11	H	0	0,09
	17	F	0	0,12	H	0	0,07
	18	F	0	0,12	H	0	0,17
III	19	F	0	0,37	F	0	0,11
350	20	F	0	0,44	F	0	0,08
Dominos	21	F	0	0,18	F	0	0,15
Amis	22	F	-0,02	-0,08	F	-0,01	0,12
	23	F	-0,01	0,21	F	0	0,14
	24	F	0	0,25	F	-0,03	0,48
	25	F	0	0,3	F	0,32	0,32

Table 4. 4. Taux de convergence calculés à partir des rapports de log-vraisemblance calculés avec Alizée. Pour chaque sujet, la première et la deuxième colonne correspondent réciproquement au taux de convergence moyen et à la déviation standard calculée sur les quatre conditions.

La Figure 4. 9 illustre les résultats obtenus avec la reconnaissance du locuteur. On observe qu'en général les taux de convergence sont plus élevés dans la condition « amis » par rapport à la condition « Inconnus » avec une divergence notable pour la paire 22. Nous avons également calculé la corrélation entre les résultats obtenus pour chaque méthode en fonction du corpus comme le montre la Table 4. 5. Les taux de corrélations obtenus sont élevés et significatifs dans le cas de la condition « amis », démontrant la nécessité d'augmenter la taille du corpus après notre expérience préliminaire.

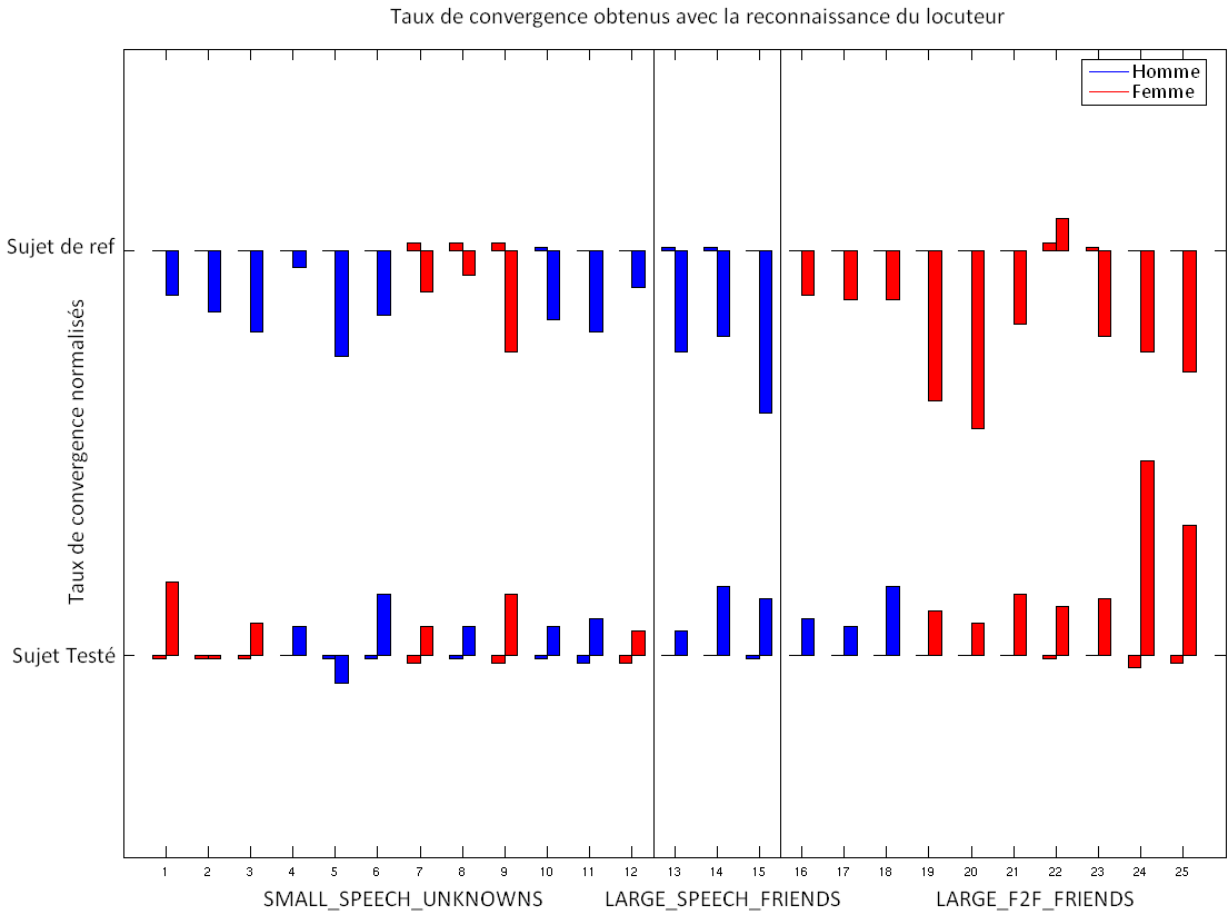


Figure 4. 9. Résultats obtenus avec la reconnaissance du locuteur. Pour chaque interaction, la première barre correspond au taux de convergence calculé sur la deuxième partie de chaque pré-test (i.e. condition de contrôle) et la deuxième barre correspond au taux de convergence calculé sur les signaux d'interaction. On observe des taux de convergence plus élevés pour des paires de même sexe et particulièrement pour des paires de femmes.

	Petit corpus	Grand corpus
Sujet de référence	0.39	0.53*
Sujet Testé	-0.04	0.54*

Table 4. 5. Scores de corrélation obtenus pour les deux méthodes proposées. Les étoiles indiquent la significativité des résultats obtenus (* = p<0.01, ** = p<0.05, * = p<0.1). Les corrélations obtenues dans le cas du corpus II sont fortes et significatives prouvant la validité de la deuxième méthode.**

4.2.4 Commentaires

La reconnaissance du locuteur nous a permis d'obtenir une mesure de convergence globale, qui ne se focalise pas uniquement sur des segments critiques. La comparaison entre les deux méthodes utilisées nous montre qu'il s'agit d'une méthode fiable. Elle pourrait permettre d'étudier le phénomène de convergence phonétique en conditions moins contrôlées à condition d'obtenir une taille de corpus assez importante.

4.3 Reconnaissance de Parole

Nous avons voulu caractériser la convergence phonétique en utilisant la reconnaissance de parole. Pour cela, nous avons utilisé des modèles de Markov cachés. Comme pour la reconnaissance du locuteur, ces modèles sont probabilistes, nous nous affranchissons ainsi des problèmes dues à la variabilité intra-locuteur et à la variabilité du canal. La démarche est similaire à celle utilisée en reconnaissance du locuteur : nous entraînons des modèles HMM de nos locuteurs grâce à leur pré-test et nous utilisons les modèles de chaque locuteur de chaque paire pour faire de la reconnaissance de parole croisée sur les signaux de pré-test et sur les signaux de l'interaction.

Nous allons d'abord décrire les modèles de Markov cachés.

4.3.1 Les modèles de Markov cachés (HMM)

Les modèles de Markov cachés peuvent être vus comme des machine à états où à chaque instant t l'automate change d'états et émet alors une observation O_t. On dit qu'ils sont « cachés » car seules les observations sont accessibles et pas la suite d'états responsable des observations. Cette observation est une réalisation d'une loi de probabilité dite « d'émission » que l'on note $b_j(O_t)$, j correspondant au nombre d'états du modèle de Markov caché. Le changement d'état dépend également d'une loi de probabilité qu'on appelle probabilité « de transition » notée a_{ij}. Si le HMM est d'ordre 1, la probabilité « de transition » de l'état i vers l'état j ne dépend pas des états antérieurs. On aura alors :

$$a_{ij} = P(E_t = j | E_{t-1} = i) \ \ avec \ \ a_{ij} > 0 \ \ et \ \ \sum_{j=1}^{N_E} a_{ij} = 1 \qquad (4.11)$$

N_E étant le nombre d'états de l'automate.

La probabilité « d'émission » b_j est généralement définie par un mélange de gaussiennes. On a donc, avec les notations du §4.2.1 :

$$b_j(o) = \sum_{i=1}^{M} \beta_i N(o, \mu_i, \Sigma_i)$$

$$avec \ N(o, \mu_i, \Sigma_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_i|}} e^{-\frac{1}{2}(o-\mu_i)^T \Sigma_i^{-1} (o-\mu_i)} \qquad (4.12)$$

Il ne reste plus qu'à définir l'état initial de l'automate. On notera $\pi_i = P(\mathrm{E}_0 = i)$, la probabilité que l'automate soit dans l'état i à l'instant $t=0$, avec $\sum_{i=1}^{N_e} \pi_i = 1$. Le HMM sera alors entièrement défini par le jeu de paramètres

$$\left\{N_E, \{\pi_i\}, \{a_{ij}\}, \{b_j\}\right\} \quad avec \quad 1 \le i \le N_E \ \ et \ \ 1 \le j \le N_E \qquad (4.13)$$

La Figure 4. 10 illustre un modèle de Markov caché à trois états E_1, E_2 et E_3. Chaque flèche et son coût a_{ij} traduisent la probabilité, pour l'automate, de passer d'un état à l'autre.

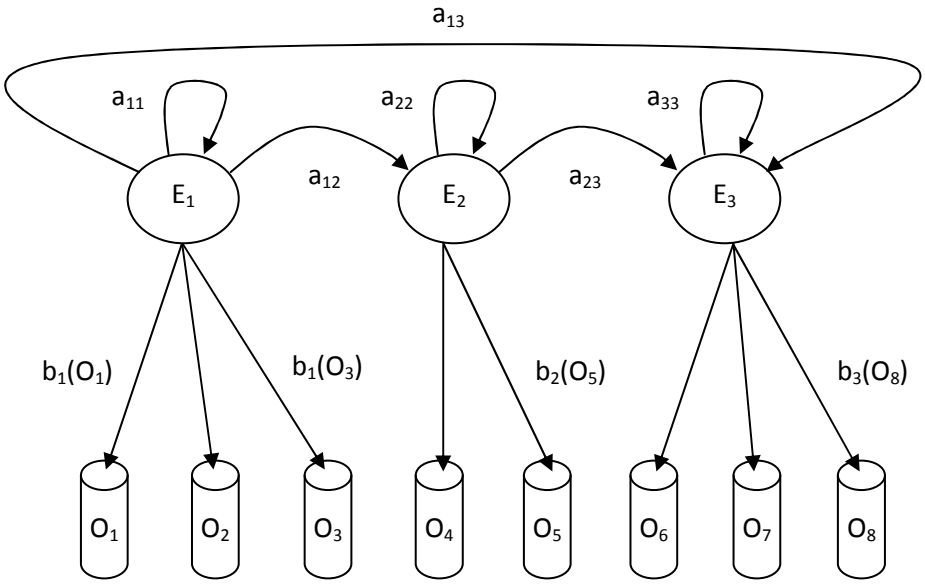


Figure 4. 10. Représentation d'un modèle de Markov caché à trois états E_1, E_2 et E_3. Les huit observations $O= [O_1, …, O_8]$ vont être générées par la séquence d'état S=1, 1, 1, 2, 2, 3, 3, 3.

En utilisant HTK, nous avons d'abord initialisé nos modèles à l'aide de l'algorithme de Viterbi, puis les paramètres des modèles ont été estimés à partir des observations en utilisant l'algorithme de *Baum-Welch* qui est une forme particulière de l'algorithme « EM » cité précédemment. Cet algorithme va estimer les paramètres itérativement de manière à maximiser la vraisemblance.

4.3.2 Méthodes

Grâce à HTK, nous avons entraîné, pour chaque locuteur, un modèle HMM de chaque phonème en contexte indépendant. Nous avons utilisé des HMMs à cinq états. Les HMMs ont été entraînés à partir des 12 premiers coefficients MFCC, de l'énergie et des deltas de ces paramètres. Nous avons ensuite utilisé les modèles de chaque interlocuteur de chaque paire pour procéder à de la reconnaissance de parole croisée. La procédure est semblable à celle utilisée pour la reconnaissance du locuteur (voir §4.2.2). Des tests t appariés ont été utilisés pour comparer les changements des distributions de scores de reconnaissance.

4.3.3 Résultats

La Figure 4. 11 représente les distributions des scores de reconnaissance pour deux paires différentes (*lmb* avec *rl* en haut et *ALa* avec *SM* en bas). On suppose que chaque phonème a été étiqueté correctement pour utiliser la reconnaissance de parole. Les phonèmes d'un locuteur ont d'abord été reconnus avec ses propres HMMs et ensuite avec les HMMs de son interlocuteur. On s'attend à obtenir des scores élevés lorsqu'on utilise la reconnaissance de parole sur les données du

pré-test d'un locuteur en utilisant son propre modèle (les modèles sont entraînés avec ses données) et des scores plus faibles lorsqu'on utilise le modèle de son interlocuteur. Le score de reconnaissance pour chaque voyelle est la log-vraisemblance moyenne par trame calculée pour l'état central du HMM correspondant. La différence entre les scores traduit la distance entre les locuteurs. La convergence sera alors caractérisée par une diminution des scores de reconnaissance lorsqu'on utilise le modèle propre du locuteur et une augmentation de ces scores lorsqu'on utilise le modèle de son interlocuteur. A gauche de la Figure 4. 11, on représente les distributions des scores de reconnaissance pendant le pré-test alors que la partie droite de la Figure 4. 11 concerne ces mêmes distributions pendant l'interaction. Les scores sont effectivement plus élevés pour les phonèmes produits par un locuteur et reconnus par ses propres HMMs. On remarque une faible adaptation des sujets. On observe bien un déplacement faible vers la gauche pour lmb_rl_lmb, rl_lmb_rl et SM_ALa_SM et sur la droite pour lmb_rl_rl, rl_lmb_lmb et SM_ALa_ALa. Cela traduit le fait que lmb, rl et SM sont moins bien reconnus par leur propre modèle pendant l'interaction et mieux reconnus par le modèle de leur interlocuteur. On peut également voir que l'amplitude du déplacement est plus importante dans le cas où les deux sujets se connaissent (interaction du bas sur la Figure 4. 11) par rapport au cas où les sujets sont deux inconnus (interaction du haut sur la Figure 4. 11).

En complément de cette caractérisation, nous avons calculé le score de reconnaissance moyen par locuteur en utilisant la reconnaissance croisée. Nous avons calculé le score de reconnaissance moyen sur les données de pré-test et d'interaction d'un locuteur en utilisant son propre modèle HMM et en utilisant ensuite le modèle HMM de son partenaire. On obtiendra alors un taux correspondant à la dégradation de la reconnaissance lorsqu'on utilise le modèle propre d'un locuteur et un taux correspondant à l'amélioration de la reconnaissance lorsqu'on utilise le modèle de son interlocuteur. On obtiendra une convergence si le taux d'amélioration est positif et le taux de dégradation négatif. On calcule le taux d'amélioration (t_a) en calculant le quotient entre la différence des scores de reconnaissance moyen obtenus sur les signaux d'interaction et de pré-test d'un locuteur en utilisant le modèle de son interlocuteur et la différence des scores de reconnaissance moyen obtenus sur les signal de pré-test d'un locuteur en utilisant son propre modèle et le modèle de son interlocuteur (voir équation 4.14). Pour le taux de dégradation (t_d), on remplace le numérateur par la différence des scores de reconnaissance moyen obtenus sur les signaux d'interaction et de pré-test d'un locuteur en utilisant son propre modèle (voir équation 4.15).

$$t_{a\ 1\to2} = \frac{(I_{l1l2})_{l2} - (P_{l1})_{l2}}{(P_{l1})_{l1} - (P_{l1})_{l2}} \tag{4.14}$$

$$t_{d\ 1\to2} = \frac{(I_{l1l2})_{l1} - (P_{l1})_{l1}}{(P_{l1})_{l1} - (P_{l1})_{l2}} \tag{4.15}$$

Le but de ce calcul est d'obtenir un taux de convergence similaire à celui obtenu en reconnaissance du locuteur (voir §4.2.2). On ne peut cependant pas trouver un taux de convergence global car la dégradation n'est pas forcément égale à l'amélioration (vu que l'entrainement d'un

modèle HMM d'un locuteur est indépendant des données de son interlocuteur). Les résultats sont regroupés dans la Table 4. 6.

			Sujet de référence						Sujet Testé				
Exp	Paires	Sexe	amélioration	dégradation	C moyen	C_{LDA}	C_{LLR}	Sexe	amélioration	dégradation	C moyen	C_{LDA}	C_{LLR}
	1	H	0,04	-0,18	0,11	0,06	0,11	F	0,1	-0,27	0,185	0,03	0,21
	2	H	0,17	-0,12	0,145	0	0,14	F	0,06	-0,01	0,035	0,02	0,05
I	3	H	0,07	-0,23	0,15	0,01	0,19	F	0,16	-0,05	0,105	0,12	0,12
186	4	H	-0,2	-0,3		0,02	0,04	H	-0,47	-0,24		0,12	0,1
Dominos	5	H	-0,34	-0,87		0,06	0,25	H	-0,17	-0,03		0,28	-0,005
médiatisé	6	H	0,08	-0,35	0,215	0,08	0,16	H	-0,14	-0,26		0,3	0,18
Inconnus	7	F	0,02	-0,03	0,025	0,02	0,06	F	-0,07	-0,24		0,15	0,13
	8	F	-0,06	0,04		0,12	0,05	H	-0,06	-0,01		0,09	0,11
	9	F	-0,03	-0,44		0,41	0,3	F	-0,07	-0,24		0,19	0,16
	10	H	0,13	-0,19	0,16	0,15	0,18	H	-0,01	-0,13		0,17	0,1
	11	H	-0,25	-0,34		0,09	0,21	H	0,24	0	0,12	0,07	0,14
	12	H	-0,04	-0,07		0,07	0,09	F	0,11	-0,03	0,07	-0,07	0,11
II	13	H	0,56	-0,07	0,315	0,68	0,21	H	-0,06	-0,09		0,24	0,22
médiatisé	14	H	0,23	0	0,115	0,32	0,24	H	0,37	-0,01	0,19	0,13	0,07
Amis	15	H	0,63	-0,3	0,465	0,4	0,44	H	0,16	-0,03	0,095	0,03	0,19
	16	F	-0,06	-0,13		0	0,13	H	0,03	-0,11	0,07	-0,03	0,18
	17	F	0,03	-0,13	0,08	0	0,12	H	0	-0,09	0,045	0,05	0,09
	18	F	-0,38	-0,21		0,06	0,13	H	0	-0,31	0,155	0,1	0,19
III	19	F	0,43	-0,82	0,625	0,14	0,38	F	0,07	0,09		0,12	0,12
350	20	F	0,07	-0,78	0,425	0,4	0,46	F	0,13	0,08		0,01	0,08
Dominos	21	F	0,16	-0,13	0,145	0,15	0,22	F	0,11	-0,22	0,165	0,22	0,15
Amis	22	F	-0,55	-0,09		0,07	-0,11	F	-0,1	-0,19		0,28	0,22
	23	F	0,09	-0,22	0,155	0,15	0,2	F	-0,08	-0,21		0,18	0,24
	24	F	-0,04	-0,43		0,22	0,23	F	0,21	-0,85	0,53	0,35	0,54
	25	F	0,09	-0,52	0,305	0,11	0,26	F	0,38	-0,36	0,37	0,15	0,28

Table 4. 6. Taux d'amélioration et de dégradation des scores de reconnaissance de parole. Pour obtenir une convergence, le taux d'amélioration doit être positif et le taux de dégradation doit être négatif. Cela traduit le fait qu'en interaction, s'il y a convergence, on se rapproche de son interlocuteur mais on s'éloigne de soi-même. Pour ces cas uniquement, nous avons calculé le taux de convergence moyen (C moyen) correspondant à la moyenne du taux d'amélioration et du taux de dégradation. On observe un seul cas de divergence très faible pour le sujet de référence 8 pour lequel le taux d'amélioration est négatif et le taux de dégradation est positif. Aucune divergence significative n'est non plus à noter par cette méthode.

On remarque que dans la plupart des cas étudiés (10 cas pour le sujet de référence et 12 cas pour le sujet testé), nous n'obtenons pas de dégradation du score de reconnaissance en utilisant le modèle d'un locuteur et une amélioration en utilisant le modèle de son interlocuteur. On ne peut donc pas conclure sur les résultats.

Nous avons cependant calculé les scores de corrélation entre les taux de convergence obtenus avec la reconnaissance de parole et ceux obtenus avec les deux autres méthodes décrites précédemment. Les Table 4. 7 et Table 4. 8 résument les résultats obtenus. Nous obtenons des scores de corrélation très élevés, en particulier avec la méthode basée sur la reconnaissance du locuteur, mais qui sont faussés par le fait que nous calculons les scores de corrélations avec les cas pour lesquels nous obtenons bien le phénomène attendu.

	Petit corpus	Grand corpus
Sujet de référence	0.38	0.28
Sujet Testé	0.6	0.85**

Table 4. 7. Scores de corrélation obtenus pour les taux de convergence moyens calculés avec la reconnaissance de parole et l'analyse discriminante linéaire. Les étoiles indiquent la significativité des résultats obtenus (* = p<0.005, ** = p<0.01, * = p<0.05).**

	Petit corpus	Grand corpus
Sujet de référence	0.06	0.82**
Sujet Testé	0.94*	0.87***

Table 4. 8. Scores de corrélation obtenus pour les taux de convergence moyens calculés avec la reconnaissance de parole et la reconnaissance du locuteur. Les étoiles indiquent la significativité des résultats obtenus (* = p<0.005, ** = p<0.01, * = p<0.05).**

4.3.4 Commentaires

Le problème majeur de la méthode décrite dans ce paragraphe est le manque de données pour entraîner nos modèles. En effet, dans la littérature, on voit qu'il faut collecter plusieurs heures de corpus pour entraîner les modèles HMMs. Les données que nous avons collectées pendant nos expériences sont donc insuffisantes pour que les résultats soient probants. Vu la tendance des résultats obtenus, on peut cependant supposer que cette méthode serait valable avec un corpus plus important. Les résultats présentés dans ce paragraphe ont été choisis de manière à illustrer la méthode, mais ils ne sont pas concluants sur l'ensemble des interactions enregistrées. De plus, nous avons calculé les scores de reconnaissance normalisés pour chaque phonème hors contexte. Nous aurions pu examiner l'état central de chaque phonème de manière à obtenir une méthode de caractérisation comparable, dans la démarche, à l'analyse discriminante linéaire.

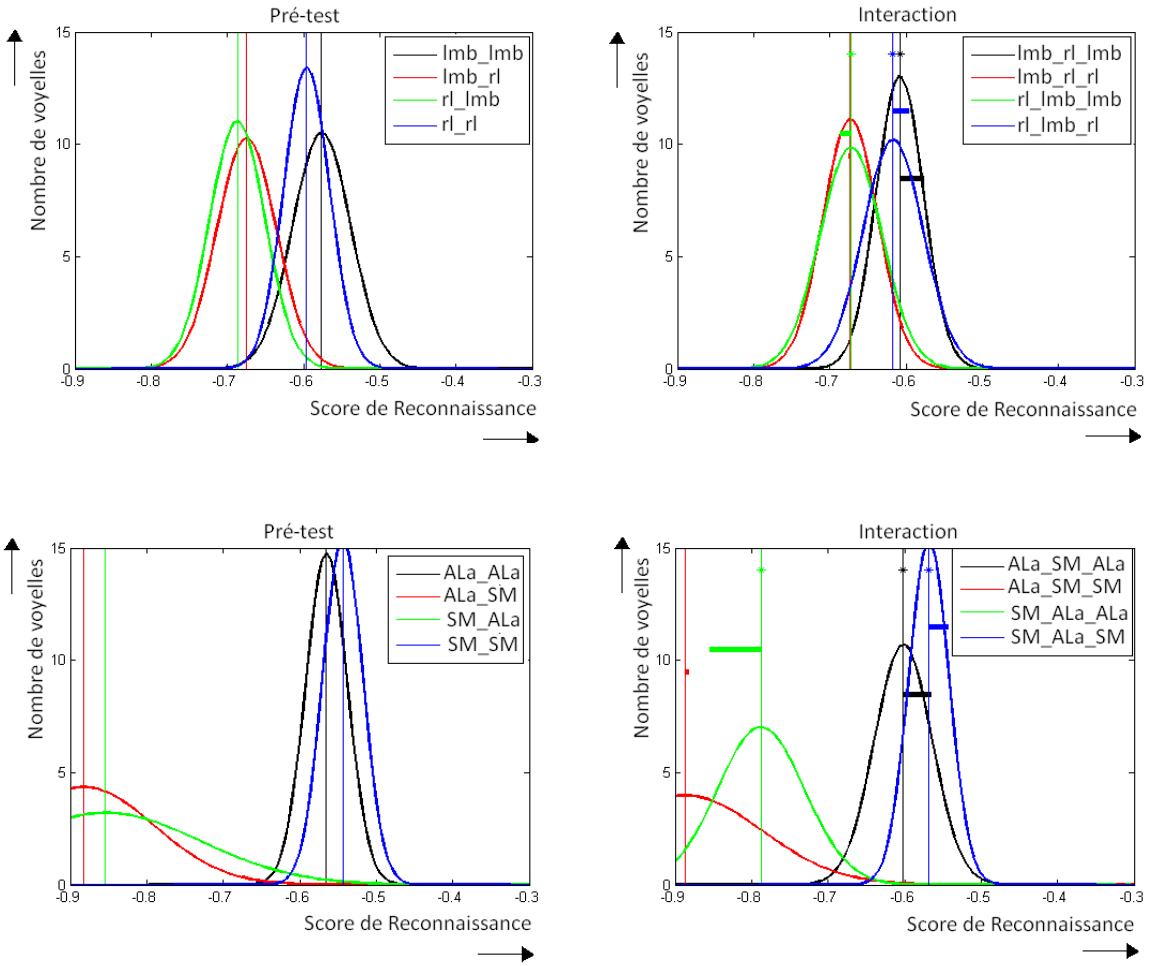


Figure 4. 11. Distribution des scores de reconnaissance pour les voyelles des mots disyllabiques produits par les locuteurs. La reconnaissance est effectuée en utilisant le modèle HMM d'un locuteur et également ce lui de son interlocuteur. On s'attend à obtenir des scores hauts en utilisant le modèle propre d'un locuteur. A gauche, on représente les scores pour les mots lus en isolation pendant le pré-test; Ces données sont utilisées pour entraîner le HMM de chaque locuteur. On remarque qu'en utilisant le modèle propre de chaque interlocuteur (lmb_lmb et rl_rl en haut, ALa_ALa et SM_SM en bas), on obtient de meilleurs scores de reconnaissance par rapport à la reconnaissance croisée (lmb_rl et rl_lmb en haut, ALa_SM et SM_ALa en bas). A droite, les mêmes mots ont été prononcés mais pendant l'interaction. Dans ce cas, on s'attend à une diminution des scores de reconnaissance en utilisant le modèle propre d'un locuteur (lmb_rl_lmb et rl_lmb_rl en haut, ALa_SM_ALa et SM_ALa_SM en bas) et à une augmentation des scores de reconnaissance quand on utilise la reconnaissance croisée (lmb_rl_rl et rl_lmb_lmb en haut, ALa_SM_SM et SM_ALa_ALa en bas). Ici, on remarque uniquement une faible adaptation (déplacement faible vers la gauche pour lmb_rl_lmb, rl_lmb_rl et SM_ALa_SM et sur la droite pour lmb_rl_rl, rl_lmb_lmb et SM_ALa_ALa). Les étoiles traduisent la significativité du déplacement calculées grâce à un test t apparié ($p<0.05$).

Après avoir développé plusieurs méthodes de caractérisation objective de la convergence phonétique, nous allons étudier les différents facteurs qui peuvent influencer l'amplitude de la convergence tels que la connaissance du contenu linguistique, les liens sociaux ou encore la fréquence lexicale, etc.

4.4 Répétition vs. Interaction

Nous supposons que la connaissance à priori du contenu linguistique va influencer la production des sujets. Pour tester cette hypothèse, nous avons demandé aux sujets de répéter le domino précédemment énoncé par leur partenaire avant de prononcer leur propre domino. Aucune consigne d'imitation n'a été donnée. Nous avons ensuite séparé automatiquement les signaux de répétition des signaux d'interaction pour étudier les taux de convergence de ces deux signaux.

En utilisant les dominos verbaux et des mots dissyllabiques, nous avons introduit un problème dû à l'accentuation des syllabes initiales et finales. En effet, les voyelles comparées pendant l'analyse discriminante linéaire ne sont pas forcément placées en même position dans le mot et donc ne sont pas accentuées de la même façon. On compare donc des patterns prosodiques qui sont différents. En demandant aux sujets de répéter, on va alors pouvoir comparer les mêmes patterns. De plus, nous supposons que si les sujets connaissent la cible linguistique à produire, ils vont être influencés par celle-ci et imiter involontairement la production de leur partenaire, on parle alors d'écholalie. Les taux de convergence obtenus devraient donc être plus importants.

La Figure 4. 12 donne les résultats obtenus avec l'analyse discriminante linéaire et la Figure 4. 13 illustre les résultats obtenus avec la reconnaissance du locuteur. Nous remarquons quelques cas pour lesquels le taux de convergence moyen est plus élevé pendant les répétitions par rapport à celui engendré par l'interaction. On observe ce phénomène pour le sujet testé des paires 1, 4, 7, 8, 10 et pour le sujet de référence de la paire 9, dans le cas de l'analyse discriminante linéaire. Pour le sujet de référence des paires 5, 7, 8 et 10 ainsi que pour le sujet testé des paires 2, 3 et 9, il n'y a pas de différence flagrante entre les deux taux de convergence comparés. Alors que pour le sujet de référence des paires 1, 2, 3, 4, 6 et le sujet testé des paires 5 et 6, on retrouve un taux de convergence plus élevé en interaction qu'en répétition. Dans ce dernier cas, on met l'accent sur le but communicatif de la convergence en interaction.

Ces résultats confirment que l'accentuation des voyelles n'influence pas le taux de convergence calculés avec l'analyse discriminante linéaire. En effet, si cela avait été le cas, on aurait obtenu des taux de convergence plus forts pour les patterns d'accentuation similaires, ainsi lorsque les sujets répétaient le domino de leur interlocuteur.

Nous avons également utilisé la méthode basée sur la reconnaissance du locuteur pour tester notre hypothèse. La Table 4. 9 reporte les résultats obtenus avec cette méthode. On observe n'aucune différence significative entre les taux de convergence calculés par les deux méthodes sur les répétitions et sur l'interaction à part pour le sujet de référence de la paire 9 et le sujet testé de la paire 4.

Exp		Paires		Sujet de référence					Sujet Testé			
				Rep		Int			Rep		Int	
	Paires		Sexe	LDA	LLR	LDA	LLR	Sexe	LDA	LLR	LDA	LLR
Répétitions Amis	1	1	F	0,03	0,31	0,12	0,37	H	**0,05**	-0,07	**0,02**	-0,05
	2	2	F	0,26	0,41	0,34	0,45	F	0,21	0,23	0,2	0,29
Répétitions Famille	3	3	F	0,18	0,73	0,29	0,73	F	0,29	0,09	0,31	0,11
	4	4	F	0,37	0,41	0,60	0,41	F	**0,34**	**0,25**	**0,26**	**0,15**
	5	5	F	0,28	0,56	0,29	0,59	F	0,09	0,23	0,16	0,24
	6	6	F	0,16	0,58	0,32	0,59	H	-0,01	0,06	0,02	0,13
	7	7	H	0,17	**0,14**	0,2	**0,12**	H	**0,30**	**0,64**	**0,16**	**0,63**
	8	8	H	0,13	**0,4**	0,12	**0,34**	F	**0,26**	0,09	**0,19**	0,1
	9	9	H	**0,27**	**0,11**	**0,25**	**0,26**	H	0,56	0,5	0,57	0,55
	10	10	H	0,10	0,22	0,11	0,31	F	**0,09**	0,02	**0,06**	0,09

Table 4. 9. Taux de convergence trouvés avec les deux méthodes développées pour les signaux de répétitions et d'interaction. Ces résultats concernent principalement des paires composées de personnes de la même famille, à part pour les deux premières paires qui servent de contrôle pour comparer des interactions entre des amis et des personnes d'une même famille. On a mis en gras les locuteurs pour lesquels le taux de convergence en répétition est plus élevé que celui en interaction. Les cases correspondantes ont été coloriées en rose pour l'analyse discriminante linéaire et en jaune pour la reconnaissance du locuteur.

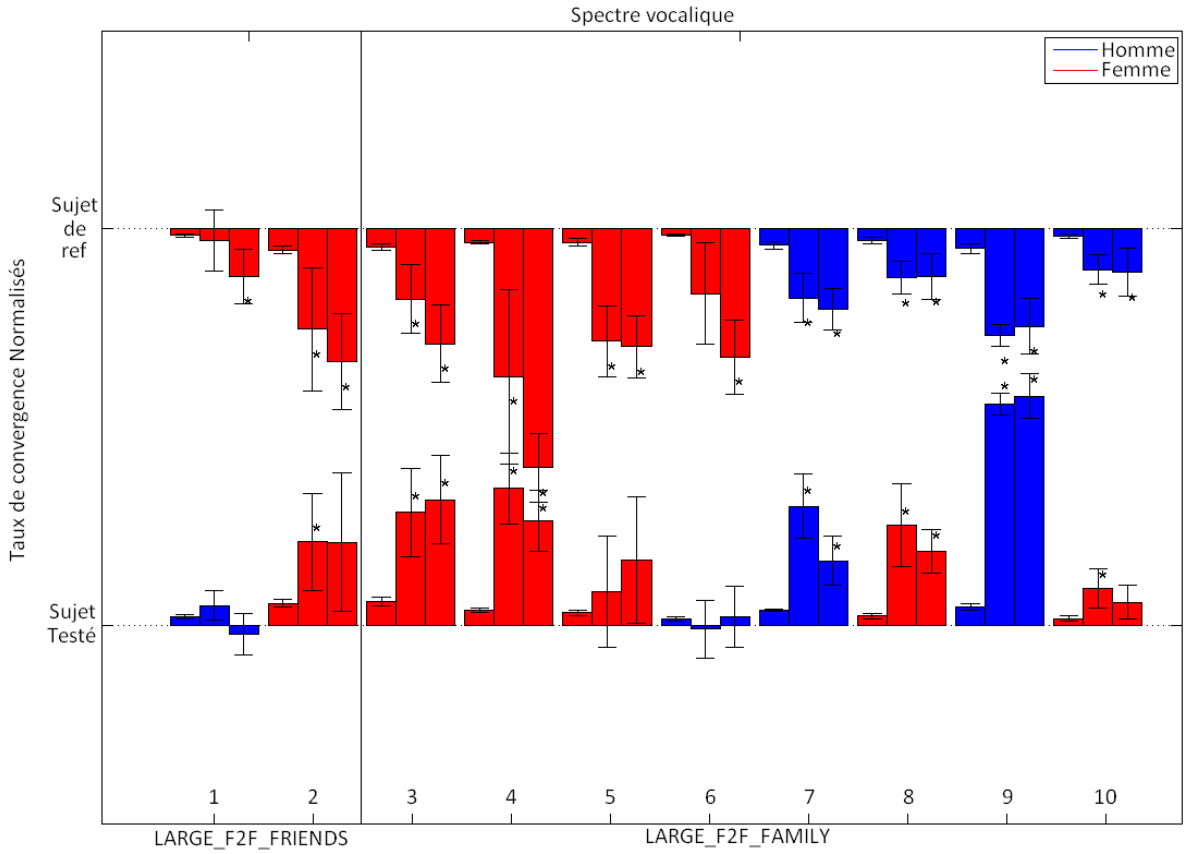


Figure 4. 12. Taux de convergence moyens sur 100 itérations calculés grâce à l'analyse discriminante linéaire sur les MFCCs des cibles vocaliques pour les signaux correspondant aux répétitions (colonne du centre) et ceux correspondant à l'interaction (colonne de droite). Pour chaque locuteur, la première colonne correspond aux pré-tests. Les taux de convergence du sujet de référence ont été inversés pour souligner le rapprochement des sujets en interaction. Une étoile indique si les distributions en interaction et en répétition sont significativement différentes par rapport au pré-test ($p<0.1$). On n'observe pas le résultat attendu, i.e. un taux de convergence plus élevé en répétition qu'en interaction, à part pour quelques paires, par exemple pour le sujet testé des paires 1, 4, 7, 8, 10 et pour le sujet de référence de la paire 9.

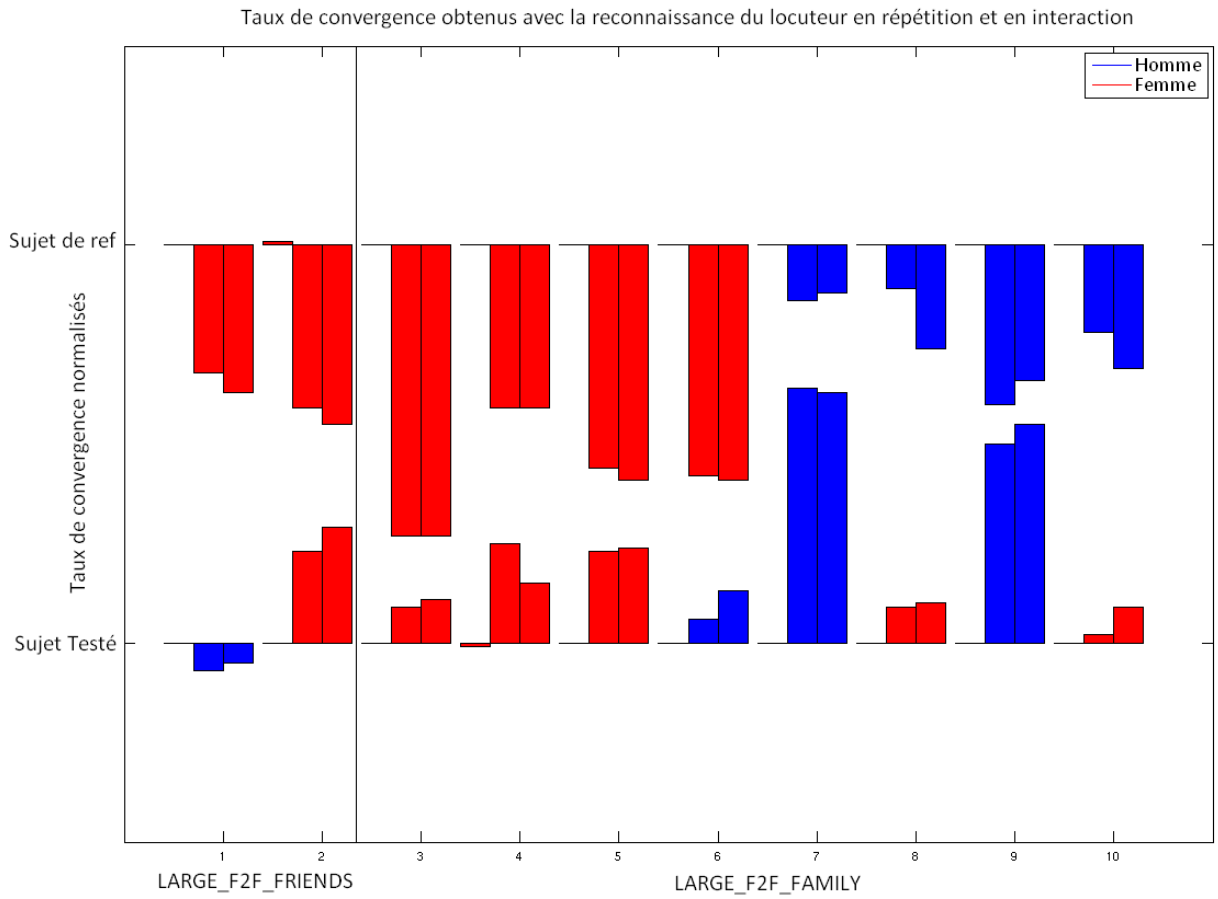


Figure 4. 13. Taux de convergence globaux obtenus à partir de la reconnaissance du locuteur pour les signaux de répétitions (colonne du milieu) et ceux d'interaction (colonne de droite). Pour chaque locuteur, la première colonne correspond aux pré-tests. On remarque que les taux de convergence en répétitions ne sont pas plus élevés que ceux en interactions à part pour les paires 7 et 9 pour le sujet de référence et 4 et 7 pour le sujet testé.

Comme précédemment, nous avons calculé les taux de corrélation des taux de convergence pour comparer les deux méthodes mais ils ne donnent aucun résultat significatif. Il faudrait donc augmenter le nombre d'interactions étudiées.

Nous avons ensuite étudié l'impact du lien social entre les sujets en testant le paradigme avec des paires provenant d'une même famille. On suppose que des pour des personnes qui vivent ensembles, les modèles internes sont plus souvent activés et donc on obtiendra une convergence plus rapide et plus importante.

4.5 Amis vs. Famille

Comme nous avons obtenu de meilleurs résultats avec des paires composées de familiers plutôt qu'avec des paires composées d'inconnus, nous avons décidé d'observer le phénomène au sein d'une même famille. Puisque les sujets se connaissent depuis très longtemps et se côtoient beaucoup, on s'attend à obtenir des taux de convergence encore plus élevés que pour les deux précédentes conditions. Nous avons enregistré deux familles différentes :

1. Pour la première famille, le sujet de référence était une fille âgée de 26 ans. Elle a interagi avec ses deux sœurs (31 et 23 ans) et ses parents (53 et 50 ans).
2. Pour la seconde famille, le sujet de référence était un garçon (25 ans) et il a interagi avec son frère (22 ans), sa sœur (19 ans) et ses parents (50 et 49 ans).

La Figure 4. 14 présente les résultats obtenus. On remarque immédiatement les interactions 4 et 9 pour lesquelles on atteint un taux de convergence égal à environ 60%. Pour la paire 4, le sujet de référence s'adapte davantage au sujet testé alors qu'on observe le phénomène inverse pour la paire 9. Ces deux interactions correspondent à des paires de sœurs (paire 4) et de frères (paire 9). Dans les deux cas, le sujet le plus jeune s'adapte à son aîné. On ne remarque pas ce phénomène pour les paires 3 et 10 qui correspondent également à des interactions au sein d'une même fratrie. Pour la même paire 10, les faibles taux de convergence s'expliquent par le fait qu'on compare deux personnes de sexe différent. Alors que pour la paire 3, on a plutôt une convergence mutuelle des deux sœurs. Si on s'intéresse maintenant aux interactions avec les parents, on remarque deux comportements différents. Dans la première famille, on remarque les parents ne change quasiment pas de comportement entre le pré-test et l'interaction, en particulier le père. Alors que dans la seconde famille, les parents s'adaptent bien à leurs fils bien que cette adaptation soit moindre par rapport à celle de la paire 9. Il est également intéressant de remarquer que les distributions des taux de convergence en interaction sont très souvent significativement différentes de celles du pré-test correspondant.

La méthode utilisant la reconnaissance du locuteur a également été utilisée. Les résultats sont reportés dans la Table 4. 9 dans la partie précédente et illustrés sur la Figure 4. 15. On confirme qu'on obtient une convergence plus forte pour les paires de même sexe. Nous n'obtenons cependant pas le même pattern de résultats. Nous observons des convergences fortes pour toutes les paires de même sexe et indépendantes des relations familiales (paires 3, 4, 9, 10 pour les fratries vs. 5, 6, 7, 8 pour les parents).

Le nombre d'interactions enregistrées ne nous permet pas de conclure pour le moment sur un pattern de convergence au sein d'une même famille. Les résultats obtenus avec la première famille nous laissent penser qu'il est possible que la convergence phonétique soit liée à un phénomène de hiérarchie dans la famille (i.e. l'autorité parentale) et soit un marqueur possible des relations de dominance au même titre que la registre ou la qualité de voix (Campbell 2004). Il faut cependant confirmer cette hypothèse par l'enregistrement d'autres familles. Nous avons cependant confirmé notre hypothèse selon laquelle des taux de convergence plus élevés pouvaient être atteint au sein d'une même famille (60% en famille vs. 40% pour des amis).

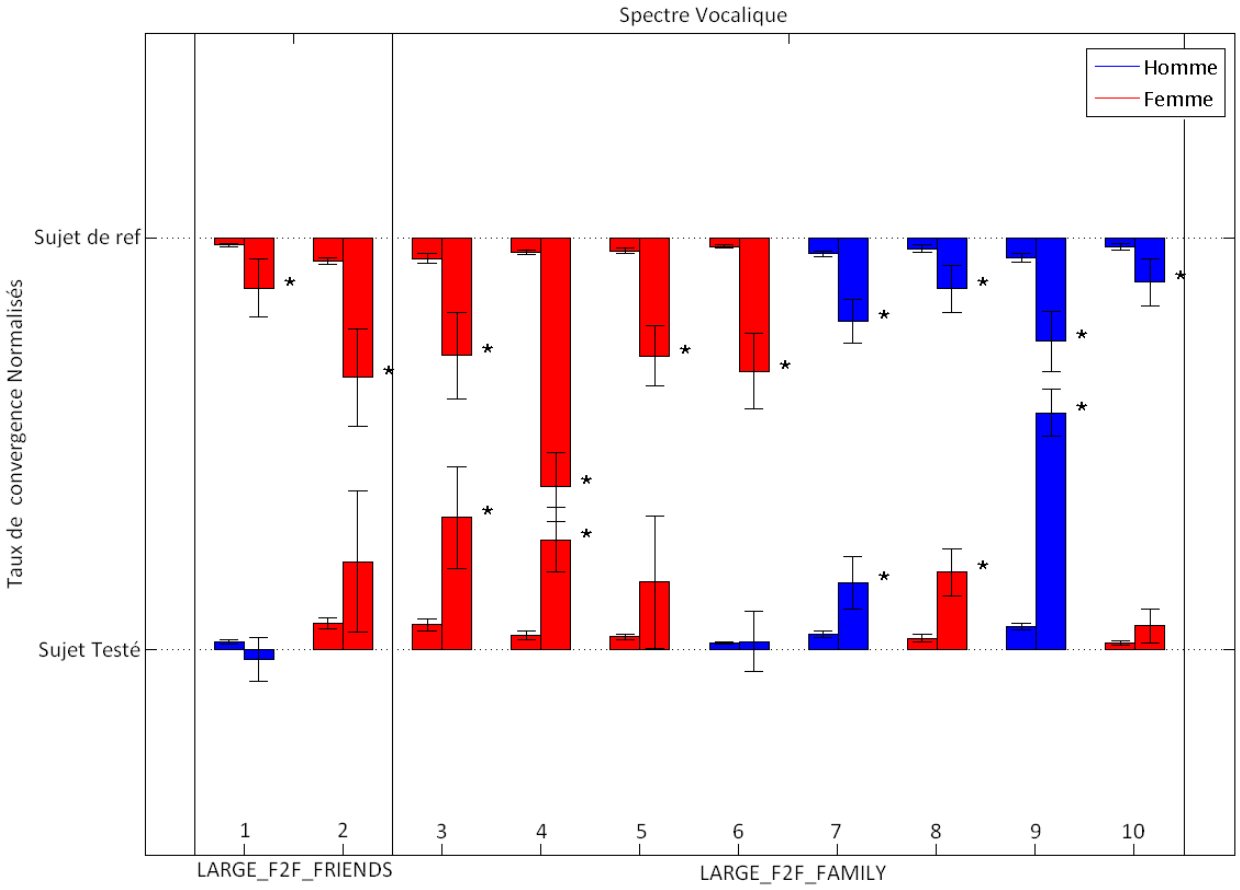


Figure 4. 14. Taux de convergence calculés grâce à une analyse discriminante linéaire sur le spectre. Les deux premières interactions sont des paires d'amis alors que les huit suivantes correspondent à des interactions entre des personnes d'une même famille. On remarque principalement deux interactions (4 et 9) pour lesquelles les taux de convergence sont très importants. Elles correspondent à des interactions entre deux sœurs pour la paire 4 et deux frères pour la paire 9. Il est intéressant de remarquer que dans la première famille (paire 3 à 6), l'adaptation du sujet testé est plus forte entre les sœurs (paires 3 et 4) plutôt qu'entre le sujet de référence et chacun de leur parent (paire s 5 et 6), les parents ne changent quasiment pas de comportement, en particulier le père. Pour la deuxième famille (paires 7 à 10) ce phénomène est moins prononcé, on retrouve une forte adaptation entre les deux frères mais pas entre le frère et la sœur qui confirme que le phénomène est facilité pour des paires de même sexe.

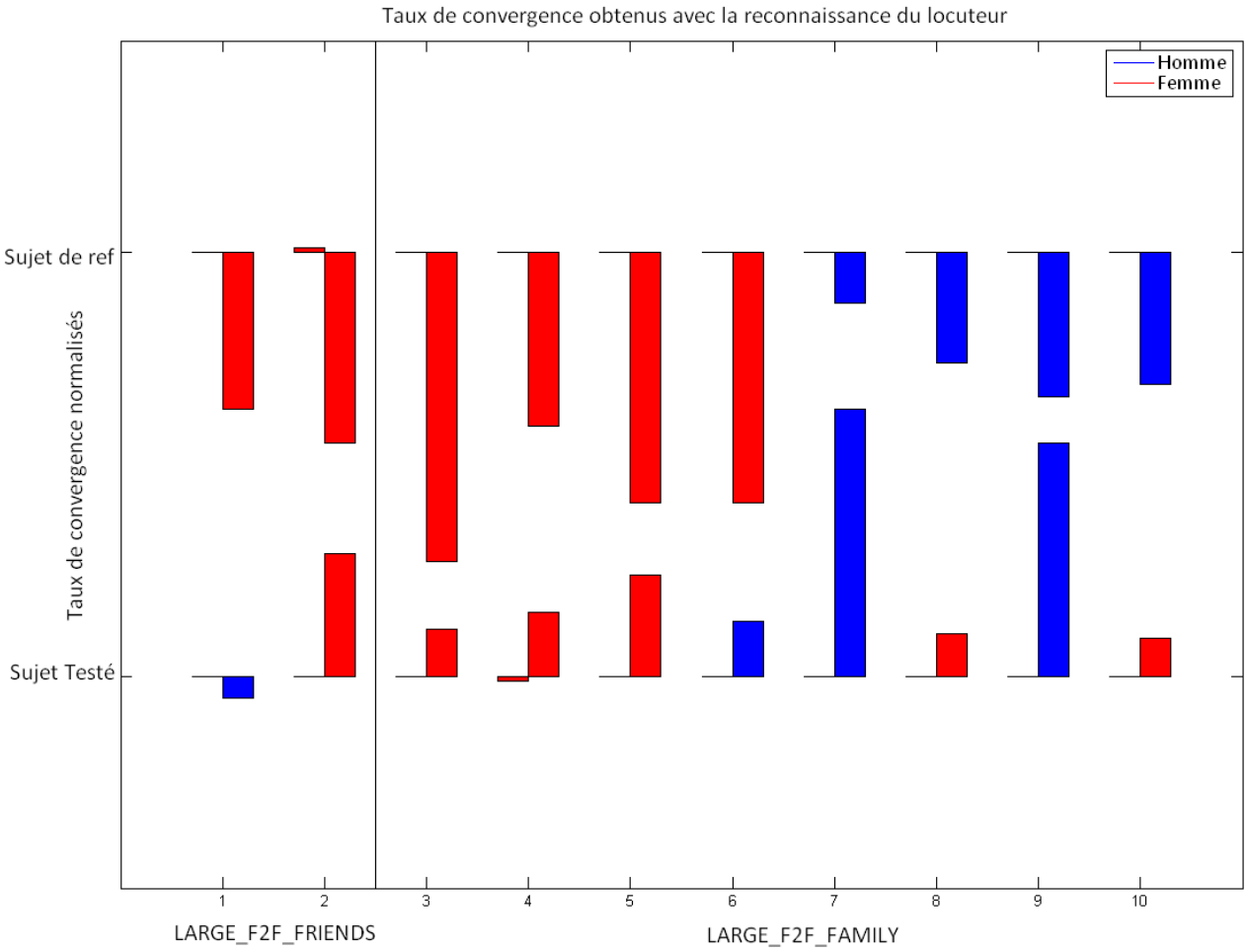


Figure 4. 15. Taux de convergence calculés grâce à la reconnaissance du locuteur. Les deux premières interactions sont des paires d'amis alors que les huit suivantes correspondent à des interactions entre des personnes d'une même famille. On remarque que les taux de convergence sont moins élevés pour des paires de sexe différent (paires 1, 6, 7 et 9). Ensuite on observe de forts taux de convergence pour le sujet de référence des paires (3, 5 et 6) et le sujet testé des paires 7 et 9. On obtient des patterns différents par rapport aux résultats obtenus avec l'analyse discriminante linéaire. Pour la première famille (paire 3 à 6), on a une convergence forte du sujet de référence quelque soit sa relation avec le sujet testé alors que pour la seconde famille, on a plutôt une convergence forte du sujet testé (voir paires 7 et 9). Les taux de convergences observées sont cependant plus forts que ceux obtenus avec des paires d'amis (voir paires 1 et 2).

Nous confirmons notre hypothèse selon laquelle les taux de convergence sont plus élevés au sein d'une famille en comparaison à des amis. Nous allons maintenant observer l'impact de la fréquence lexicale des mots utilisés sur l'amplitude de la convergence.

4.6 Convergence et Fréquence lexicale

Comme nous l'avons présenté au Chapitre 2, les mots ont été choisis de manière à pouvoir étudier l'impact de la fréquence lexicale sur le taux de convergence. D'après Goldinger (1998), les mots ayant une faible fréquence lexicale vont être davantage imités que ceux avec une fréquence lexicale élevée. Pour étudier ce phénomène, nous avons séparé les mots prononcés pendant l'interaction en deux groupes de même taille en fonction de leur fréquence lexicale (voir la Figure 2. 8). Nous avons ensuite calculé les taux de convergence moyens en fonction des groupes étudiés. Les fréquences lexicales ont

été obtenues à l'aide de la base de données LEXIQUE (http://www.lexique.org/). Le premier groupe correspond aux mots qui ont une fréquence lexicale inférieure à 1. Cette étude a été faite sur le corpus comportant 350 dominos afin de comparer des groupes de taille conséquente.

Interaction	Sujet de référence		Sujet testé	
	f faible	f élevée	f faible	f élevée
1 (F/H)	**0,04**	**-0,01**	-0,02	-0,03
2 (F/H)	**0,02**	**-0,01**	**0,05**	**0,02**
3 (F/H)	**0,08**	**0,07**	0,08	0,12
4 (F/F)	0,12	0,14	0,06	0,15
5 (F/F)	0,42	0,45	**0,04**	**-0,02**
6 (F/F)	0,16	0,18	0,16	0,22
7 (F/F)	**0,15**	**0,00**	**0,29**	**0,26**
8 (F/F)	0,06	0,17	0,16	0,20
9 (F/F)	**0,25**	**0,09**	0,30	0,40
10 (F/F)	0,07	0,12	0,07	0,19
11 (F/H)	0,08	0,13	-0,05	-0,03
12 (F/F)	**0,33**	**0,32**	0,17	0,23
13 (F/F)	0,23	0,35	**0,30**	**0,29**
14 (F/F)	0,56	0,60	0,23	0,30
15 (F/F)	0,27	0,27	0,10	0,14
16 (F/H)	0,28	0,37	**-0,02**	**0,00**
17 (H/H)	0,17	0,21	**0,20**	**0,15**
18 (H/F)	0,09	0,11	**0,17**	**0,15**
19 (H/H)	**0,26**	**0,20**	**0,61**	**0,56**
20 (H/F)	0,07	0,13	**0,07**	**0,04**

Table 4. 10. Taux de convergence moyens en fonction de la fréquence lexicale des mots prononcés par chaque sujet pendant l'interaction. Les mots sont séparés en deux groupes, le premier groupe comporte des mots de fréquences lexicales faibles et le second des mots de fréquences lexicales élevées. On a mis en gras les interactions pour lesquelles on a obtenu le résultat attendu (i.e. un taux de convergence plus élevé pour une fréquence lexicale plus faible), la tendance reste cependant très faible.

La Table 4. 10 et la Figure 4. 16 présentent les résultats obtenus. On remarque quelques cas pour lesquels le taux de convergence moyen est plus élevé pour les mots de fréquences lexicales faibles. Ce phénomène est retrouvé pour les paires 1, 2, 3, 7, 9, 12 et 19 pour le sujet de référence et 2, 5, 13, 17, 18, 19, 20 pour le sujet testé. La différence entre les taux de convergence moyen des deux groupes reste cependant très faible (en moyenne 0.05). On trouve également l'effet inverse soit un taux de convergence moyen plus élevé pour une fréquence lexicale plus élevée pour les paires 8, 10, 11, 13, 16 pour le sujet de référence et 4, 6, 9, 10, 12, 14 et 16 pour le sujet testé. Une nouvelle fois on ne remarque pas de phénomène émergent, ce qui souligne le manque de données pour pouvoir conclure sur nos résultats.

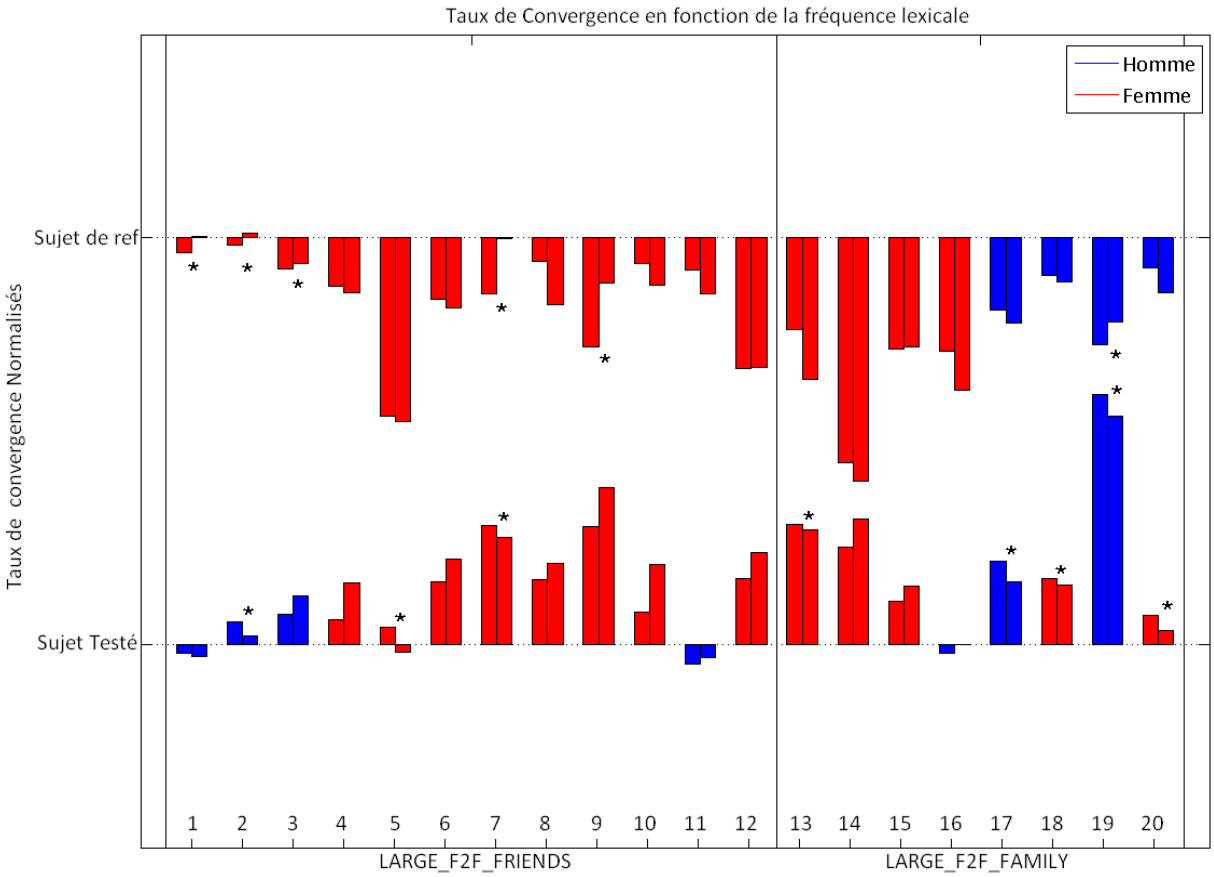


Figure 4. 16. Taux de convergence calculés grâce à une analyse discriminante linéaire sur le spectre pour les deux groupes de fréquences lexicales différentes. La colonne de gauche correspond au groupe de mots avec une fréquence lexicale faible alors que la colonne de droite correspond à celui dont les mots ont une fréquence lexicale élevée. On s'attend donc à ce que les colonnes de droite montrent des taux de convergence plus élevés. Les cas correspondant à cette conclusion sont marqués par une étoile. On remarque qu'ils ne sont pas systématique et que l'amplitude du phénomène est très variable.

On va maintenant observer l'impact du temps sur la convergence phonétique.

4.7 Evolution de la convergence en fonction du temps

Dans la littérature, on retrouve différents résultats concernant le lien entre l'amplitude de la convergence et le temps. Par exemple, Kousidis (2008) observe une convergence immédiate qui n'évolue pas au cours de l'interaction alors qu'Aubanel (2011) et Delvaux et Soquet (2007) ont observé que la convergence augmentait au fur et à mesure des phases de leur paradigme. Nous allons donc observer l'évolution des taux de convergence calculés grâce à l'analyse discriminante linéaire.

Pour cela, nous avons calculé les coefficients de la régression linéaire entre les taux de convergence obtenus pour chaque item et la place de l'item en question dans l'interaction. On rappelle que les taux de convergence des sujets testés sont centrés autour de 0 et ceux des sujets de référence autour de 1 de manière à illustrer le rapprochement des deux sujets en interaction. Ainsi l'amplitude de la convergence augmentera avec le temps si le coefficient directeur de la droite de régression linéaire est positif pour le sujet testé (il part de 0 et se rapproche de 1) et négatif pour le sujet de

référence (il part de 1 et se rapproche de 0. La Table 4. 11 nous donne les coefficients directeurs obtenus pour chaque interaction.

Interaction	Condition	Sujet de référence	Sujet testé
		Pente	Pente
1 (H/F)	Inconnus Corpus I	**-7,71E-04**	-1,85E-03
2 (H/F)		**-9,54E-04**	-2,75E-04
3 (H/F)		**-5,25E-04**	**1,98E-03**
4 (H/H)		**-8,63E-04**	-4,84E-05
5 (H/H)		1,78E-03	**4,51E-04**
6 (H/H)		1,58E-03	-6,51E-04
7 (F/F)		2,96E-03	**3,79E-03**
8 (F/H)		**-1,41E-03**	**6,47E-04**
9 (F/F)		2,39E-03	-8,76E-04
10 (H/H)		1,53E-03	**9,58E-04**
11 (H/H)		1,06E-03	-1,34E-03
12 (H/F)		**-1,03E-03**	-8,03E-05
13 (H/H)	Amis Corpus II	1,07E-03	-6,72E-04
14 (H/H)		2,96E-04	**8,81E-04**
15 (H/H)		**-2,88E-04**	-4,07E-04
16 (F/H)		**-2,85E-04**	-2,65E-04
17 (F/H)		**-8,04E-04**	-3,26E-04
18 (F/H)		7,01E-05	**7,66E-04**
19 (F/F)		4,26E-04	-6,28E-04
20 (F/F)		**-2,08E-03**	**4,67E-04**
21 (F/F)		1,27E-04	-5,24E-04
22 (F/F)		**-5,51E-05**	-6,06E-04
23 (F/F)		1,47E-03	-1,35E-04
24 (F/F)		**-2,68E-04**	**1,67E-04**
25 (F/F)		4,34E-04	-3,14E-04
26 (F/H)	Famille Corpus II	5,77E-04	**7,80E-05**
27 (F/F)		6,43E-05	**5,39E-04**
28 (F/F)		**-1,38E-04**	**2,04E-04**
29 (F/F)		**-1,10E-04**	-5,67E-04
30 (F/F)		5,39E-04	-1,20E-03
31 (F/H)		4,32E-04	-9,17E-04
32 (H/H)		**-3,00E-05**	-5,08E-04
33 (H/F)		1,70E-04	-4,87E-04
34 (H/H)		**-3,57E-05**	**2,96E-04**
35 (H/F)		**-9,28E-04**	-6,41E-05

Table 4. 11. Coefficients directeurs des droites de régression linéaire des taux de convergence en fonction du temps. Les taux de convergence augmentent avec le temps si le coefficient directeur est négatif pour le sujet de référence et positif pour le sujet testé. Ces cas ont été mis en évidence en gras dans le tableau.

On observe 17 cas pour lesquelles la convergence augmente au fur et à mesure de l'interaction pour les sujets de référence et 13 cas pour les sujets testés. Les coefficients directeurs obtenus sont cependant très faible (de l'ordre de 10^{-5} à 10^{-3}).

Nous étudions maintenant l'apport de la convergence pour l'interaction en observant le lien entre les taux de convergence et la performance (mesurer ici en termes de temps de réponse) pour accomplir la tâche.

4.8 Convergence et performance

Nous avons enfin voulu étudier le lien entre la convergence et la performance d'une interaction. Pour cela nous avons défini le temps de tour de parole (TTT pour Turn Taking Time) comme le délai entre le début de la dernière voyelle d'un domino prononcé par un locuteur et le début de la première voyelle du domino suivant prononcé par son interlocuteur. La Figure 4. 17 montre le lien entre la convergence et l'évolution du TTT pendant les interactions: pour des taux de convergence moyens, le degré de convergence du sujet de référence vers son interlocuteur est corrélé avec des TTT de plus en plus courts de son partenaire (r = -0.77). En d'autres termes, plus le sujet de référence va s'adapter à son interlocuteur et plus les TTT de celui-ci vont être courts, le sujet testé accélère le rythme de l'interaction si son partenaire s'adapte à lui. Nous ne trouvons pas cet effet pour les sujets testés, nous obtenons un coefficient de corrélation de 0.26 (voir Figure 4. 18). Cela confirme que le rôle et l'expérience de chaque participant sont des paramètres importants qui vont influencer leur comportement et leur performance (Pardo, 2010).

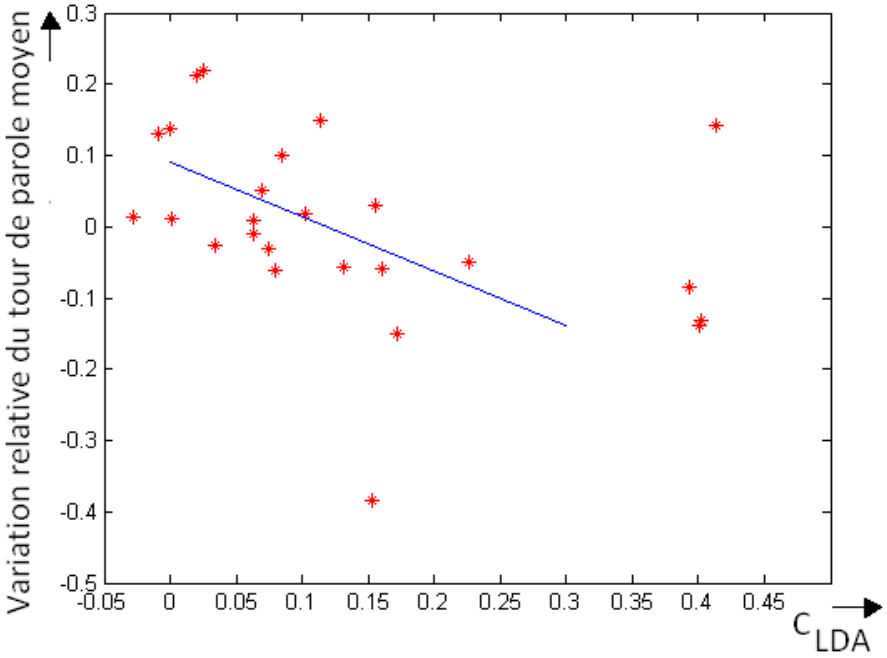


Figure 4. 17. Variation relative du tour de parole moyen du sujet testé en fonction du taux de convergence de son partenaire C_{LDA}. On remarque que plus le sujet de référence s'adapte à son interlocuteur, plus celui-ci va accélérer le rythme de l'interaction.

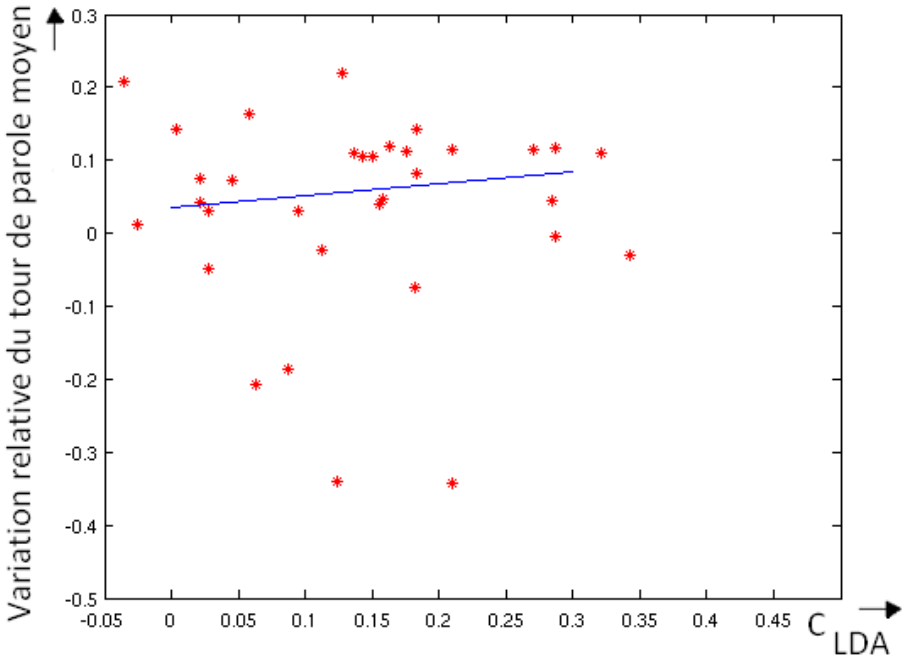


Figure 4. 18. Variation relative du tour de parole moyen du sujet de référence en fonction du taux de convergence de son partenaire C_{LDA}. On ne remarque pas de lien entre le taux de convergence du sujet testé et le temps de réponse du sujet de référence.

4.9 Prosodie

Gregory a étudié la convergence du f0 en conversation libre en étudiant les coefficients de corrélation entre les f0 des partenaires au début, au milieu et à la fin de la conversation. Il a remarqué que ces coefficients augmentaient au cours de l'interaction, ainsi le lien entre les fréquences fondamentales des deux protagonistes était de plus en plus fort. Levitan (2011) a également étudié la convergence de paramètres prosodiques tels que la fréquence fondamentale ou l'intensité. Pour cela, il a divisé son corpus en deux parties de deux manières différentes. Il a d'abord séparé le premier jeu puis la session complète pour étudier la convergence à deux échelles de temps différentes. Il a comparé les différences de moyenne et de valeur maximale de chaque paramètre entre un sujet et son interlocuteur et conclut qu'il y a une convergence si les différences sont plus faibles dans la seconde partie étudiée. Il observe des convergences qui ne sont pas systématiquement significatives, en particulier pour la fréquence fondamentale. Les résultats confirment cependant une convergence à plus long terme.

4.9.1 Méthode

Pour étudier la convergence du f0 et de la durée des phonèmes, nous avons fait une régression linéaire de ces paramètres en fonction du temps pour chaque interlocuteur, nous étudierons ainsi l'évolution de la convergence en fonction du temps comme l'ont fait Gregory ou Levitan. Nous avons ensuite calculé le point d'intersection des deux droites de régression grâce à l'équation suivante :

$$\text{Si} \quad y_1(t) = a_1 t + b_1 \text{ et } y_2(t) = a_2 t + b_2$$
$$\text{alors} \quad t_{int} = \frac{b_2 - b_1}{a_1 - a_2} \tag{4.16}$$

Nous avons ensuite divisé l'abscisse obtenue pour le point d'intersection par le nombre d'items qui ont été prononcés de manière à normaliser le résultat obtenu. Comme on peut le voir sur la Figure 4. 19, si t_{int} est positif alors on a un rapprochement des paramètres étudiés au cours du temps, il y a donc convergence entre les deux sujets, alors que si t_{int} est négatif on a une divergence des paramètres. Les coefficients directeurs des pentes de régression vont également nous donner une information importante. En effet, elles traduisent à quel point les sujets vont converger/diverger. Une dissymétrie des pentes nous indiquera une dissymétrie au niveau du pattern de convergence.

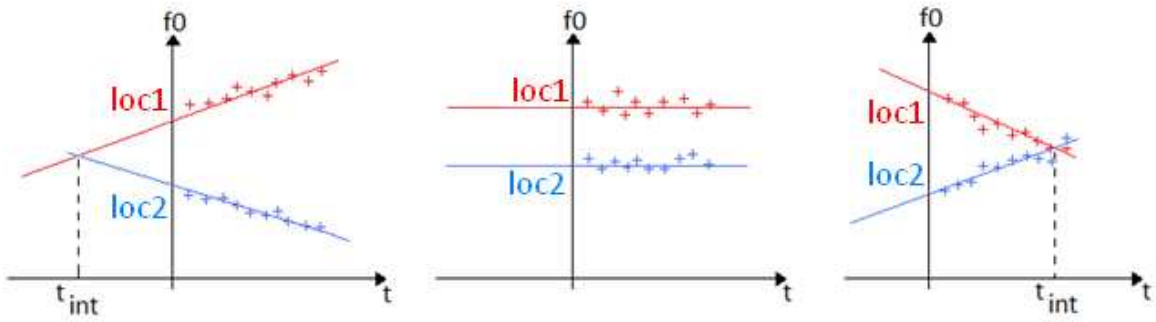


Figure 4. 19. Méthode utilisée pour étudier la convergence prosodique, à gauche nous avons un exemple de divergence pour lequel le point d'intersection des régressions linéaires est négatif ; un exemple de stationnarité au centre, dans ce cas, en théorie les droites ne se croisent jamais, en pratique la valeur absolue de t_{int} sera très grande, et un exemple de convergence à droite, plus t_{int} sera proche de zéro et plus les droites convergeront vite.

Nous avons donc comparé les points d'intersection obtenus avec les pré-tests des deux sujets et ceux obtenus pendant leur interaction. Plusieurs cas sont alors possibles (voir Equation (4.17)):

1. Si le point d'intersection des pré-tests (tp_{int}) a une abscisse négative, on a donc une divergence des paramètres étudiés pendant les pré-tests des deux sujets. On aura alors une convergence en interaction si le point d'intersection ti_{int} est supérieur à tp_{int} sinon les paramètres divergeront par rapport aux pré-tests.
2. Si le point d'intersection des pré-tests (tp_{int}) a une abscisse positive – i.e. les paramètres étudiés ont naturellement tendance à se rapprocher – alors on aura convergence si ti_{int} est positif et inférieur à tp_{int} – les paramètres se rapprochent plus vite en interaction – sinon on aura une divergence.

En résumé on remarquera une convergence des paramètres des sujets pendant les interactions si :

$$\begin{cases} tp_{int} < 0 \quad \text{et} \quad ti_{int} > tp_{int} \\ \text{ou} \\ tp_{int} > 0 \quad \text{et} \quad 0 < ti_{int} < tp_{int} \end{cases} \tag{4.17}$$

4.9.2 Résultats

Il faut noter que plus le point d'intersection des régressions linéaires sera positif et proche de zéro, plus les paramètres auront tendance à converger rapidement. Réciproquement, plus il sera éloigné positif et éloigné de zéro et plus les paramètres convergeront lentement. La Table 4. 12 regroupe les résultats obtenus pour le f0 et la durée des voyelles. La convergence n'est pas systématique pour les deux paramètres étudiés. Pour le f0, on observe 16 cas de convergence qui ne sont pas dépendantes du sexe des sujets étudiés. On obtient 19 cas d'adaptation pour la durée des voyelles étudiées. L'adaptation des sujets entre eux peut s'établir très rapidement comme pour les paires 5, 20, 28 et 32 pour le f0 et les paires 3, 4, 6, 12, 17, 20, 24, 31, 32 et 34 pour la durée des voyelles. Comme on a divisé le coefficient obtenu par le nombre d'items, on suppose que t_{int} doit être inférieur à 1 pour que l'alignement se fasse pendant l'interaction sinon cela signifie juste que l'adaptation est plus lente. On observe également quelque cas de divergence pour le f0 et la durée des voyelles. Nous avons comparé ces résultats avec les taux de convergence. En général, les cas de convergences des paramètres prosodiques correspondent à des cas de convergence phonétique. Seules les paires 12 et 17 ne suivent pas cette tendance.

Les résultats obtenus démontrent également que la convergence de la fréquence fondamentale est très lente. En effet, les valeurs obtenues pour l'abscisse du point d'intersection des régressions linéaires est très souvent supérieur à 1 (qui correspondrait à une convergence totale). Cela rejoint les conclusions de Levitan.

4.9.3 Commentaires

Nous avons utilisé plusieurs méthodes pour essayer de caractériser l'adaptation de la fréquence fondamentale moyenne et de la durée des voyelles. Nous avons analysé la corrélation entre les fréquences fondamentales des sujets pendant leur pré-test et pendant l'interaction et également la variation moyenne de la différence des fréquences fondamentales pendant les pré-tests des sujets et pendant l'interaction. Aucune de ces méthodes n'a été concluante car elles examinaient peut-être le phénomène de manière trop globale. Cette méthode nous a permis d'observer des cas de convergence au niveau prosodique mais elle ne nous permet pas de savoir comment ce phénomène est régi pendant l'interaction. S'agira-t-il d'une adaptation mutuelle ou y aura-t-il une dominance de l'un des deux sujets ?

f0				d			
Interaction	t_{int} en Pré-test	t_{int} en Interaction	Convergence	Interaction	t_{int} en Pré-test	t_{int} en Interaction	Convergence
1 (h/f)	-9,1825	2,28	**oui**	1 (h/f)	8,9276	7,6307	**oui**
2 (h/f)	70,8951	16,4688	**oui**	2 (h/f)	1,2227	1,5135	non
3 (h/f)	47,8952	9,9611	**oui**	3 (h/f)	0,3168	0,1496	**oui**
4 (h/h)	5,8675	7,8829	non	4 (h/h)	0,7357	0,0698	**oui**
5 (h/h)	9,9063	0,8822	**oui**	5 (h/h)	2,8022	-6,6306	non
6 (h/h)	1,2024	42,1334	non	6 (h/h)	0,1865	0,194	non
7 (f/f)	1,7004	-1,1868	non	7 (f/f)	-0,4136	0,7375	**oui**
8 (f/h)	85,4776	23,9128	**oui**	8 (f/h)	-1,7568	28,8999	**oui**
9 (f/f)	-0,6532	1,4726	**oui**	9 (f/f)	0,7702	2,4654	non
10 (h/h)	4,7303	-2,2254	non	10 (h/h)	-13,177	-5,8193	**oui**
11 (h/h)	2,7994	-7,3188	non	11 (h/h)	-2,9061	-1,3413	**oui**
12 (h/f)	42,5408	-67,506	non	12 (h/f)	-1,0234	0,907	**oui**
13 (h/h)	1,9306	5,0133	non	13 (h/h)	1,6304	-1,3796	non
14 (h/h)	3,2042	-0,7335	non	14 (h/h)	-1,2442	2,2799	**oui**
15 (h/h)	2,4802	-0,9799	non	15 (h/h)	-2,9717	84,1167	**oui**
16 (f/h)	-22,1913	29,6736	**oui**	16 (f/h)	0,9977	-0,1889	non
17 (f/h)	83,6862	8,146	**oui**	17 (f/h)	1,8993	0,8189	**oui**
18 (f/h)	52,1732	7,2607	**oui**	18 (f/h)	-0,8048	4,0498	**oui**
19 (f/f)	-2,814	2,241	**oui**	19 (f/f)	5,0761	32,8098	non
20 (f/f)	0,1803	0,8422	non	20 (f/f)	-0,0948	0,6072	**oui**
21 (f/f)	4,8653	-12,4303	non	21 (f/f)	-0,7221	1,3474	**oui**
22 (f/f)	-3,1199	0,1554	**oui**	22 (f/f)	0,0494	-1,1299	non
23 (f/f)	-2,8961	-4,2322	non	23 (f/f)	1,0514	-0,4298	non
24 (f/f)	2,3138	12,6595	non	24 (f/f)	0,5838	0,1375	**oui**
25 (f/f)	-60,4701	6,4268	**oui**	25 (f/f)	2,6503	-0,7195	non
26 (f/h)	183,4696	-63,2029	non	26 (f/h)	0,5431	1,6387	non
27 (f/f)	0,0821	1,2225	non	27 (f/f)	-0,3357	-0,4093	non
28 (f/f)	1,1021	0,1326	**oui**	28 (f/f)	0,9169	1,0612	non
29 (f/f)	4,1533	-0,4419	non	29 (f/f)	0,033	-0,2287	non
30 (f/f)	-0,3355	1,0035	**oui**	30 (f/f)	-1,0888	12,4553	**oui**
31 (f/h)	-25,6159	-110,4335	non	31 (f/h)	-0,5083	0,5117	**oui**
32 (h/h)	0,9592	0,5673	**oui**	32 (h/h)	-5,1784	0,1065	**oui**
33 (h/f)	-3,6216	868,7901	**oui**	33 (h/f)	1,7907	-1,6894	non
34 (h/h)	0,6235	-9,044	non	34 (h/h)	0,8919	0,698	**oui**
35 (h/f)	6,3186	50,9928	non	35 (h/f)	0,8723	50,9928	non

Table 4. 12. Abscisses des points d'intersection des régressions linéaires de f0 (à gauche) et de la durée des voyelles (à droite) en fonction du temps pour chaque paire de sujets pendant leur pré-test et l'interaction. On précise le sexe de chaque sujet dans la colonne « Interaction » et on a mis en évidence en gras les paires pour lesquelles il y avait interaction. On remarque que la convergence n'est pas systématique.

4.10 Conclusions

Le paradigme des dominos verbaux nous a permis d'étudier le phénomène de convergence phonétique sous différentes conditions.

Nous avons d'abord développé une première méthode de caractérisation en utilisant l'analyse discriminante linéaire. Cette analyse nous a permis de trouver l'espace dans lequel les espaces phonétiques de référence des sujets étaient les plus séparés et nous avons ensuite projeté les données d'interaction de nos sujets dans cet espace.

Grâce à cette méthode, nous avons d'abord étudié l'adaptation entre deux inconnus en face-à-face médiatisé en faisant varier également le sexe des sujets. Nous avons alors observé une amplitude d'adaptation faible mais qui était également modulée par le sexe, en effet on obtenait une

convergence plus forte pour des paires de femmes ; ce qui est conforme aux résultats trouvés par Namy *et al.* (2002).

En accord avec ses conclusions, nous avons ensuite enregistré des paires de femmes amies ou collègues en face-à-face réel. Nous avons bien observé des taux de convergence plus élevés que pour des paires d'inconnus. En effet, les sujets ont déjà des modèles internes de leur interlocuteur, l'adaptation est donc plus rapide et plus facile. Pour confirmer cette tendance, nous avons comparé les comportements au sein d'une famille où l'on s'attend à des taux de convergence encore plus élevés puisque l'expérience commune aux sujets est plus importante. Les taux de convergence obtenus (60%) confirment nos conclusions. Nous avons également observé des comportements différents suivant les familles étudiées et la composition des paires (fratrie vs. parent/enfant). Il est donc intéressant de poursuivre ces enregistrements in situ pour étudier l'empreinte des relations sociales sur l'accommodation verbale (Gravano *et al.* 2011).

Nous avons ensuite étudié l'impact de la connaissance du contenu linguistique sur la production d'un locuteur. Nous avons supposé qu'en demandant aux sujets de répéter le domino précédemment prononcé par leur partenaire, nous allions obtenir des taux de convergence plus élevés pour les répétitions. En majorité, nous n'avons pas observé ce phénomène. Cela est peut être dû à la charge cognitive trop importante lorsqu'on demande aux sujets de répéter le domino de leur partenaire avant de prononcer le leur, ils sont davantage concentrés sur le mot qu'ils doivent choisir pour atteindre le but de la tâche.

Nous avons également comparé l'impact de la fréquence lexicale sur le taux de convergence. D'après Goldinger (1998), les sujets vont inconsciemment imiter les mots de fréquences lexicales faibles car ce sont ceux pour lesquels les sujets ont moins de représentations internes. Les résultats obtenus sont très variables, on observe le phénomène attendu uniquement pour quelques paires. Il faut cependant procéder à de nouveaux enregistrements pour confirmer la tendance.

L'inconvénient de la méthode basée sur l'analyse discriminante linéaire est qu'il faut étiqueter correctement tous les signaux enregistré. Bien que nous ayons utilisé la reconnaissance de parole (i.e. Modèle de Markov Cachés) pour segmenter automatiquement nos signaux, l'alignement a demandé beaucoup de temps. Nous avons donc développé une deuxième méthode pour caractériser la convergence phonétique basée sur la reconnaissance du locuteur.

Cette méthode ne demande aucune segmentation a priori de nos signaux. Nous entraînons un modèle global (GMM) de nos sujets à l'aide de la plateforme Alizée. Puis nous comparons les rapports de log-vraisemblance croisés pour déterminer les taux de convergence de nos sujets.

Les résultats obtenus ont validé la méthode utilisée. En effet, la corrélation des taux de convergence trouvés avec les deux méthodes est significativement élevée. Cette méthode va permettre d'utiliser des scénarios moins contrôlés pour caractériser la convergence phonétique. Elle sera également d'une importance cruciale pour doter des agents conversationnels animés de la stratégie d'adaptation phonétique.

Nous avons observé que, dans l'ensemble de la littérature traitant la convergence phonétique, l'amplitude du phénomène était faible. Si les taux de convergence que nous obtenons sont plus

amples, cela peut s'expliquer par une tâche finalisée liant de manière implicite performance et degré d'accommodation.

On peut se demander si les sujets perçoivent cette stratégie d'adaptation de leur interlocuteur, si les signatures objectives des convergences observées peuvent être effectivement perçues par le système perceptif ou si le mode de perception « phonologique » écarte ces dimensions paralinguistiques de son champ attentionnel lorsque le système perceptif est centré sur une opération de décodage lexical. Nous allons donc procéder à quelques tests pour étudier la perception de la convergence phonétique qu'elle soit en réelle interaction ou créée artificiellement à l'aide d'outils de synthèse vocale.

Chapitre 5 Perception de la convergence phonétique

Les patrons de convergence mis en exergue par les mesures objectives de rapprochement de caractéristiques phonétiques, aussi diverses que le VOT des plosives, les distributions moyennes des formants des voyelles ou des paramètres prosodiques, valident l'hypothèse d'une boucle sensori-motrice impliquant un rapprochement des représentations phonétiques de la parole. Les études montrent par ailleurs que divers facteurs liés aux interlocuteurs, à leurs relations sociales préexistantes, leurs rôles respectifs dans l'interaction (Pardo *et al.,* 2010) ainsi que le contenu linguistique et paralinguistique des informations échangées modulent cette accommodation phonétique.

La plupart des études montrent cependant que si cette dynamique est motivée et participe bien au bon fonctionnement de l'interaction, elle est largement régulée de manière inconsciente. Nous avons montré que cette dynamique est effectivement produite et dépend non seulement de la situation mais bien des caractéristiques phonétiques de l'espace sonore de l'interlocuteur. Il reste à démontrer que cette dynamique est bien perçue – même inconsciemment – par l'interlocuteur ou toute autre personne potentiellement intéressée par l'échange verbal.

La perception de la convergence a fait l'objet de quelques études phare, dont nous allons discuter les paramètres expérimentaux. Nous allons ensuite proposer un nouveau paradigme, appelé suivi de locuteur.

5.1 Paradigmes expérimentaux en perception de la convergence phonétique

Nous allons d'abord présenter différents types de test de perception que nous pourrions utiliser pour tester la sensibilité perceptive de nos sujets pour la convergence phonétique.

Les tests de discriminations sont utilisés pour tester la capacité de différentier des stimuli, plusieurs stimuli sont donc présentés pour chaque condition. Ces tests vont fournir comme résultats des taux de réponses correctes ou incorrectes pour chaque sujet. Il existe plusieurs types de tests de discrimination.

5.1.1 Test AX

Le test *AX* est très utilisé dans le domaine de la parole. A chaque tour, deux stimuli sont présentés aux sujets. Ils sont séparés par un laps de temps que l'on appelle intervalle inter stimuli (ISI). Ils sont présentés par paire de stimuli identiques ou différents et les sujets doivent identifier si les stimuli sont les mêmes ou non. Par exemple, si on a deux stimuli <A> et <B>, les paires présentées vont être <AA>, <AB>, <BA> et <BB>. En général, il y a le même nombre de paires identiques et de paires différentes pendant un test (car les sujets vont s'attendre à cette distribution) et les grandeurs mesurées sont le pourcentage de réponses correctes et le temps de réaction. L'avantage de ce test est qu'il est très simple d'expliquer la consigne aux sujets et de calculer le temps de réaction (à partir du deuxième

stimulus). Il présente cependant quelques inconvénients. Par exemple, si la tâche est trop compliquée, les sujets vont avoir à répondre que les stimuli sont toujours identiques.

Nous pourrions utiliser ce design en testant les mêmes mots mais provenant d'interactions différentes. Cela est facilement réalisable puisque nos sujets de référence ont interagi avec différents sujets. On pourrait alors testé si les sujets testés détectent le changement d'interaction. Le phénomène que nous voulons tester est cependant tellement fin que nous prendrions le risque que les sujets trouvent les stimuli constamment identiques.

Il existe également une version de ce test pendant laquelle on demande aux sujets de répondre le plus rapidement possible, ils doivent alors baser leur décision sur les traces contenues dans leur mémoire à court terme. Ce test souligne qu'il y a deux modes différents de perception : le mode auditif (qui traite le signal à un niveau détaillé mais pour lequel les traces en mémoire sont maintenues peu de temps) et le mode phonétique qui considère une représentation catégorielle des stimuli. On demande donc aux sujets de répondre rapidement (au moins en dessous de 800 ms) pour diminuer la charge mémorielle. Il faut également que l'intervalle inter stimuli soit suffisamment court, la valeur commune est 100 ms. Cette condition de rapidité peut cependant compliquer la tâche pour les sujets.

5.1.2 Test AXB

Pour le test AXB (Pardo 2006; Pardo *et al.,* 2012), ABX ou XAB (Kim *et al.,* 2011), on présente trois stimuli aux sujets et ils doivent décider si le stimulus *X* ressemble plus au stimulus *A* ou au stimulus *B.* On utilise donc deux intervalles inter stimuli qui sont habituellement égaux. La consigne est un peu plus compliquée pour les sujets mais, en présentant deux stimuli, on permet aux sujets de comparer les deux possibilités ce qui va éliminer des biais qui pouvaient être présents avec un test AX. Ce test va cependant introduire un problème lié à la mémoire, ce qui se traduit par une tendance à choisir la réponse *B*. On compense alors ce phénomène en inversant la place de *A* et *B* pour obtenir le même nombre de cas possibles.

5.1.3 Choix forcé à deux alternatives

Pendant ce test, à chaque tour, on présente deux stimuli aux sujets et on leur demande de déterminer l'ordre des stimuli. L'instruction habituelle est de la forme « Quel stimulus vient le premier entre *A* et *B* ». Ce test minimise les biais qui pouvaient être présent avec les tests précédents. Il faut cependant définir explicitement ce que signifie le terme « ordre » aux sujets ce qui fait que la consigne peut être compliquée, on peut donc utiliser des phases d'entraînements.

Comme plusieurs mots ont été prononcés en début et en fin d'interaction, nous pourrions utiliser ce design pour observer dans un premier temps si la convergence a évolué en fonction du temps et si les sujets peuvent percevoir cela. Il faudra cependant faire attention à familiariser nos sujets avec les deux voix de l'interaction pour qu'ils puissent avoir une base de critères pour juger d'un rapprochement ou non.

5.1.4 Changement de catégorie

Ce design a été créé pour être une version analogue d'une tâche de perception pour les enfants pendant laquelle l'enfant entend un flux de syllabes et on détecte sa sensibilité au changement par des indices comportementaux supposés corrélés à l'intérêt porté par l'enfant à la « nouveauté » du stimulus. Les corrélats utilisés sont le taux de succion (Eimas, 1985) ou le changement de centre d'attention par mouvement de tête (cf. *head-turn preference procedure*) ou de regard (McCartney and Panneton ,2005). Dans la version adulte, le flux de syllabes est présenté et les sujets doivent presser un bouton quand ils détectent une répétition (Toro *et al.,* 2005) ou un changement (Content *et al.,* 2001 ; Dehaene-Lambertz *et al.,* 2005). On mesure alors le nombre de changements détectés.

5.1.5 Test Oui-non

Pendant les tâches d'identification, les sujets entendent un ou plusieurs sont et on leur demande de les catégoriser. Il faut donc définir les différentes catégories.

Le design le plus simple est celui pendant lequel on demande aux sujets de compare un stimulus avec un autre mais uniquement un stimulus est présenté à chaque tour. Par exemple, on peut demander aux sujets si le stimulus étaient présent ou pas (lorsqu'on test la fréquence et le niveau de décibel par exemple) ou si le stimulus était *A* ou *B.* Ce test est très simple à expliquer aux sujets et les mesures de temps de réaction sont simples. Mais comme les sujets n'ont pas accès aux stimuli à comparer à chaque tour, la comparaison peut s'avérer difficile.

Nous pourrions utiliser ce design de test en faisant écouter aux sujets deux mots consécutifs de la chaîne de domino et en leur demandant si ils proviennent d'une vraie interaction ou de deux interactions différentes. Il faudrait pour cela que les conditions d'enregistrements de chaque interaction soit scrupuleusement identiques pour ne pas influencer la réponse des sujets.

Nous allons maintenant présenter Les études qui ont déjà utilisé des tests de perception pour caractériser la convergence phonétique.

5.2 Etudes utilisant les tests de perception

Pardo (2006) a utilisé une tâche collaboration (voir §2.1.3) pour étudier le phénomène de convergence phonétique. Cette tâche lui a permis de récupérer plusieurs répétitions de mêmes mots – nom des illustrations sur la carte utilisée – prononcés au cours de l'interaction. Ces mêmes mots ont été enregistrés en isolation pendant un pré-test pour récupérer des prononciations de référence des sujets et pendant un post-test pour étudier le phénomène d' « after-effect ». Les enregistrements des pré-tests ont également été utilisés pour appareiller les sujets, les auteurs ont choisi de créer des paires dont les f0 étaient approximativement les mêmes. Pendant l'interaction, chaque rôle (donneur/receveur) a été assigné puis les sujets ont dû reproduire cinq cartes différentes. Ils étaient séparés par une cloison pour éviter que les cartes ne soient visibles à l'autre sujet ainsi les sujets ne pouvaient pas se voir. Six hommes et six femmes ont participé à cette expérience.

Pardo a ensuite utilisé un test de perception pour juger de la similarité entre les stimuli des différentes conditions et ainsi caractériser la convergence. Pour cela elle a mis en place un test AXB inspiré de celui utilisé par Goldinger (1998). Il a été construit de la manière suivante (voir Figure 5. 1).

Le signal à juger était composé de trois répétitions d'un même label. Le label du milieu (X) correspondait à la production d'un sujet pendant la tâche et les premier et dernier labels correspondaient au même mot prononcé par l'autre sujet pendant la tâche, pendant le pré-test ou pendant le post-test. Les juges devaient alors dire si le signal du milieu ressemblait plus au premier signal (A) ou au dernier (B) en se concentrant sur la façon dont étaient articulées les consonnes et les voyelles pour obtenir une homogénéité dans les résultats.

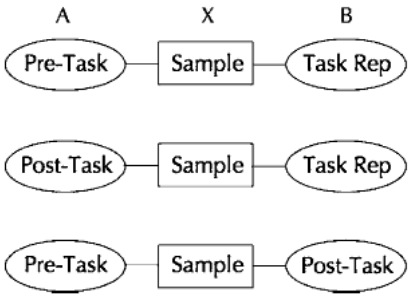


Figure 5. 1. Test AXB mis en place par Pardo pour caractériser la convergence

24 mots répétés par l'un et l'autre des sujets ont été utilisés pour construire le test de perception. 30 personnes ont participé au test AXB pour juger de la similarité entre les signaux. Le score utilisé pour mesurer le phénomène de convergence était le pourcentage de gens qui ont répondu que le signal extrait de la tâche (Sample) était plus proche du signal extrait de la tâche et répété par l'interlocuteur (Task Rep) plutôt que de son pré-test ou de son post-test ou qui ont répondu que le signal de l'interaction était plus proche du post-test plutôt que du pré-test.

Pardo a utilisé des tests statistiques ANOVA pour tester les effets du temps (tôt vs tard dans l'interaction), de la persistance (pré, tâche et post), du rôle (donneur vs receveur), du sexe.

Comme le montre la Figure 5. 2, des résultats significatifs ont été trouvés sur le rôle et le sexe. L'auteur a également trouvé une interaction entre ces deux facteurs. En effet, il semblerait que ceux qui donnaient les instructions convergeaient plus que les receveurs pour les hommes (environ 70% de similarité) et on retrouvait le pattern inverse pour les couples de femmes (environ 60% de similarité). La Figure 5. 2 montre également que les résultats étaient très dépendants des paires étudiées mais que les résultats restaient cohérents et particulièrement pour les femmes. Pardo a également montré que la convergence augmentait pendant l'interaction.

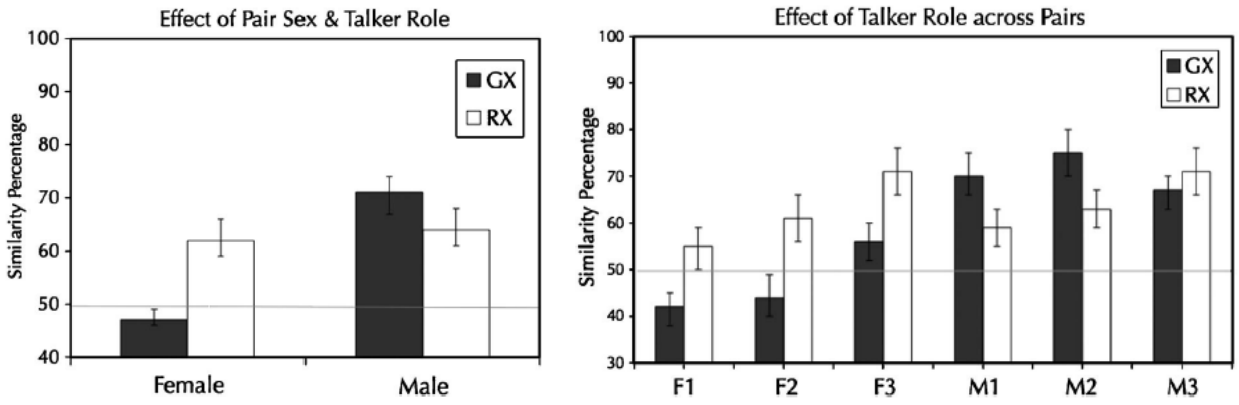


Figure 5. 2. Résultats du test AXB mené par Pardo. La barre noire correspond au taux de convergence du donneur vers le receveur et la barre blanche correspond au taux de convergence du receveur vers le donneur

Ces résultats n'étaient pas ceux attendus par l'auteur. En effet Namy *et al.,* (2002) ont obtenu des résultats complètement contradictoires par rapport à ceux trouvés par Pardo. Pendant leur expérience, ils ont montré que les femmes qui devaient imiter un modèle convergeaient plus que les hommes qui accomplissaient la même tâche et qu'il était plus facile pour les femmes de détecter une convergence. Cela rejoignait les conclusions précédentes expliquant que les femmes étaient plus habiles pour des tâches de détection de locuteurs (Nygaard and Queen, 2000; Namy *et al.,* 2002). Si les femmes étaient plus sensibles aux paramètres indexicaux des locuteurs, elles devraient alors converger plus facilement vers leur interlocuteur, ce qui contredit les résultats trouvés par Pardo. L'auteur attribue ces différences de résultats à un problème d'attention qui pouvait être différente entre une tâche d'imitation et une tâche conversationnelle.

Kim *et al.* (2011) ont étudié le phénomène d'adaptation entre l'Anglais et le Coréen mais en faisant varier la distance dialectale (voir §2.1.3). Ils ont enregistré des conversations entre 2 Anglais natifs, 2 Coréen natifs qui avaient ou non le même dialecte puis entre des Anglais natifs et non natifs. Pour caractériser la convergence, ils ont utilisé un test XAB où des sujets devaient juger de la similarité entre X et A et X et B. X était un échantillon de parole d'un locuteur (en fin d'interaction) et A et B étaient deux échantillons de son interlocuteur prélevés soit au début, soit à la fin de leur conversation. Les résultats ont montré que la convergence était plus importante pour les paires (environ 60% de similarité) qui parlaient le même dialecte que pour les autres. Les auteurs expliquaient ce phénomène par le fait que la convergence était affectée par le besoin d'intelligibilité pendant les interactions et que la charge cognitive était trop importante pendant les interactions en langue non native.

Afin de valider les méthodes de caractérisation subjectives de la convergence phonétique, nous avons synthétisé la convergence afin d'obtenir des stimuli calibrés. La comparaison des résultats obtenus en synthèse adaptative et en réelle interaction avec un taux de convergence similaire nous permettra de valider nos tests de perception. Nous avons donc choisi d'explorer une méthode de synthèse basée sur la modélisation « Harmonique plus bruit » utilisée par Hueber (Hueber, 2009).

5.3 Synthèse adaptative

5.3.1 Modélisation « Harmonique plus Bruit »

La modélisation *« Harmonique plus bruit* » plus connu sous l'acronyme « *HNM* » pour « *Harmonic plus Noise Mod*el » a été introduite par Stylianou (Stylianou*, 1990*). Nous avons utilisé cette méthode car elle permet des modifications sur la prosodie et le spectre tout en maintenant une synthèse qui semble naturelle. Cette modélisation suppose que le signal de parole peut se décomposer en deux parties : une première partie dite *harmonique* qui modélise les structures quasi-périodiques du signal et une seconde partie dite *bruitée* qui correspond aux composantes apériodiques du signal comme les bruites de friction. Ces deux composantes correspondent à *H(t)* et *B(t)* dans l'équation 4.2. Plus de détails sont donnés dans l'article de Stylianou (Stylianou, 1990) et la thèse de Hueber (Hueber, 2009).

$$s(t) = H(t) + B(t) = \sum_{j=1}^{N} (A_j \cos(2\pi j f_0 t) + \varphi_j) + (N_{gauss} * F(t)) \quad (5.1)$$

Où *N* est le nombre d'harmoniques incluent dans *H(t)*, f_0 est la fréquence fondamentale estimée, *Aj* correspond à l'amplitude au temps t de la j-ème harmonique, *Ngauss* est un bruit gaussien et *F(t)* un filtre autorégressif. Nous utilisons ici 12 composantes harmoniques et un modèle autorégressif pour la partie bruitée d'ordre 16.

La fréquence maximale de voisement f_m (déterminée à partir de l'analyse d'irrégularités dans le spectre) va déterminer la partie harmonique (avant f_m) et la partie bruitée (après f_m). Après avoir déterminé les coefficients *Aj* correspondant à l'amplitude au temps d'analyse (définis par l'équation 4.3) de la j-ème harmonique (à l'aide d'une approche par minimisation au sens des moindres carrés), on peut obtenir la partie *bruitée* de notre signal. Les coefficients LSF (*Line Spectrum Frequencies*, Itakura, 1975) ainsi que le gain (obtenu à partir de l'estimation de la variance du signal aux instants d'analyse $1/f_0$) sont obtenus à partir d'une modélisation autorégressive.

$$t_a = i \times \left(\frac{1}{f_0}\right) \; pour \; i = \left[1; \frac{f_0}{f_m}\right] \quad (5.2)$$

5.3.2 Synthèse avec la modélisation « Harmonique plus bruit »

Une interpolation entre les paramètres du locuteur *source* et ceux du locuteur *cible* nous permet d'obtenir des tailles de vecteur de paramètres semblables. Nous avons choisi d'interpoler nos paramètres sur un chemin d'alignement moyen entre les deux signaux afin de s'affranchir des différences de prosodie ou d'élocution. Nous avons ensuite utilisé l'équation 5.3 pour obtenir les nouveaux coefficients LSF et gain pour les parties harmoniques et bruitées pour procéder à la synthèse adaptative.

$$\theta_{adpt} = (1 - \alpha)\theta_A + \alpha\theta_B \quad (5.3)$$

Où θ_{adpt} correspond aux paramètres adaptés, θ_A et θ_B correspondent respectivement aux paramètres du locuteur A et du locuteur B.

Nous avons utilisé la segmentation pour récupérer les instants précis de début et fin de phonème pour procéder à une adaptation par phonème.
Nous avons synthétisé des stimuli entre 0 et 100% d'adaptation du locuteur A vers le locuteur B et réciproquement. Nous avons remarqué un souffle important sur les stimuli de synthèse obtenus, nous avons donc diminué les gains sur les parties des signaux avant le premier phonème et après le dernier phonème de chaque stimulus. Nous avons choisi pour nos tests de perception de sélectionner les stimuli à 20% d'adaptation car cela correspond au taux de convergence moyen obtenus dans le chapitre précédent (voir Figure 5. 3).

Nous avons ensuite utilisé ces stimuli pour mettre en place nos tests de perception. Deux types de tests différents ont été utilisés : des tests AXB pour confronter nos résultats à ceux de la littérature et des tests de changement de locuteur pour tester la sensibilité catégorielle des sujets.

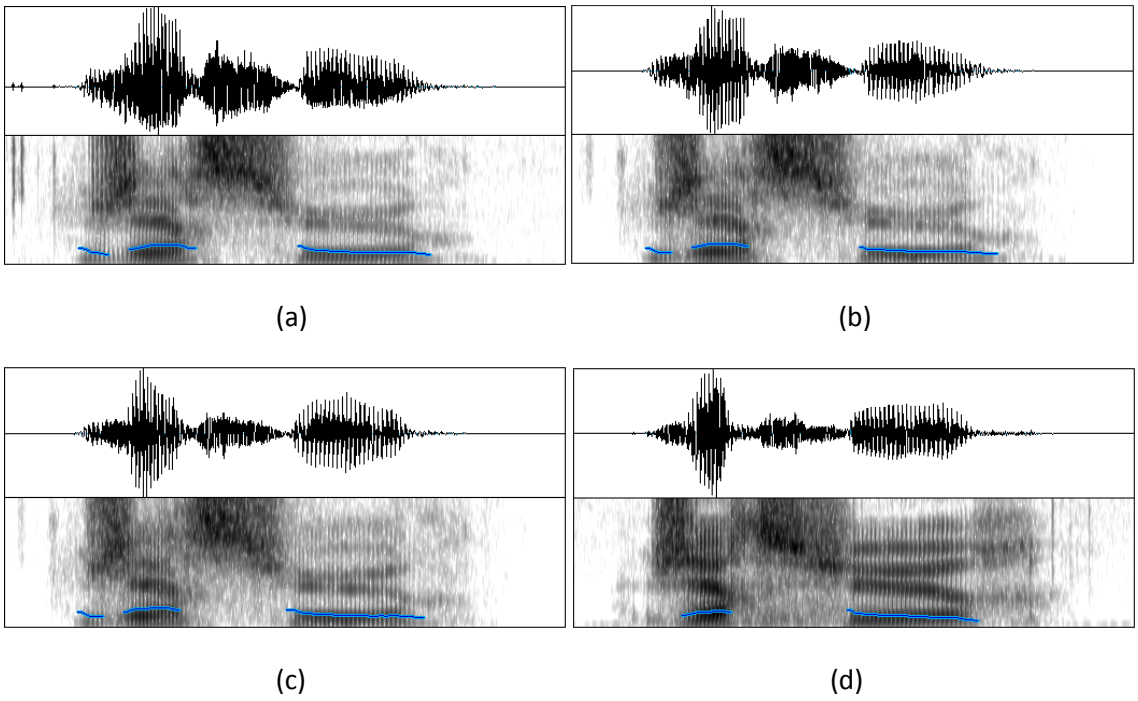


Figure 5. 3. Forme d'ondes et spectrogrammes du stimulus « gerçure » prononcé par le locuteur 1 pendant son pré-test (a), obtenu à partir d'une synthèse à 0% du locuteur 1 utilisant le modèle HNM (b), obtenu à partir d'une synthèse à 20% du locuteur 1 vers le locuteur 2 en utilisant le modèle HNM (c), et enfin prononcé par le locuteur 2 pendant son pré-test (d). La forme (c) correspond bien à une forme intermédiaire entre les formes (a) et (d).

5.4 Tests de perception

Plusieurs tests ont été mis en place pour tester la capacité des sujets à détecter l'adaptation de leur interlocuteur. On suppose que si les sujets testés peuvent percevoir l'adaptation en test de perception alors les sujets en interaction sont également capables de détecter l'adaptation de leur interlocuteur et donc de décrypter la stratégie communicative de leur interlocuteur.

5.4.1 En interaction

5.4.1.1 Test AXB

Plusieurs tests AXB ont été mis en place pour tester la perception de la convergence phonétique en interaction. Nous avons d'abord testé si les sujets pouvaient détecter la convergence en fonction du temps. Comme plusieurs stimuli étaient répété pendant l'interaction, nous avons construit notre test AXB de la manière suivante : A et B étaient des stimuli du premier locuteur prononcé soit en début soi en fin d'interaction et X correspondait au même stimulus prononcé par le deuxième interlocuteur. Le test était composé de 192. Nous avons construit un autre test AXB similaire. Comme nos sujets de référence ont interagi avec différentes personnes, nous avons testé si les sujets pouvaient détecter la convergence phonétique en comparant des signaux provenant d'une interaction réelle et d'une « fausse interaction ». Nos stimuli A et B sont des signaux prononcés par le premier locuteur soit pendant son interaction avec le deuxième locuteur soit lors d'une autre interaction. Le signal X correspond au même signal prononcé par le deuxième locuteur. Les Figure 5. 4 et Figure 5. 5 résument les designs utilisés.

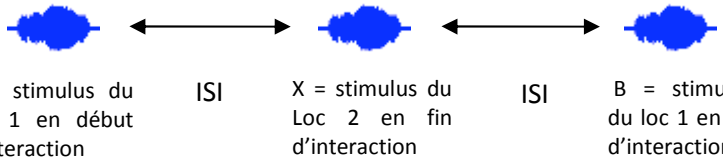


Figure 5. 4. Design du test AXB pour étudier la perception de la convergence phonétique en fonction du temps. Ici l'ISI vaut 300 ms.

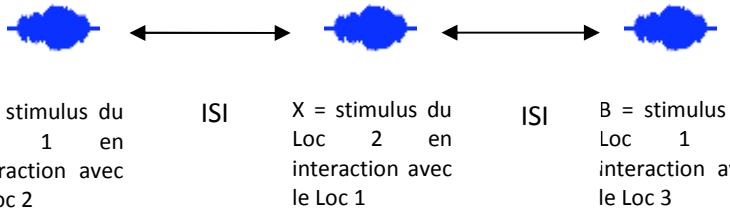


Figure 5. 5. Design du test AXB pour détecter si la convergence phonétique va influencer la perception d'une « vraie » interaction versus une « fausse » interaction. Ici l'ISI vaut 300 ms.

Nous avons fais en sorte que A et B proviennent de la même condition d'enregistrement (en interaction) et nous avons inversé A et B pour éviter tout biais. Les résultats obtenus ne nous permettent pas de conclure car nous obtenons un taux de détection de la convergence d'environ 50% ce qui correspond à de la chance dans un test avec deux choix possibles (voir Figure 5. 6).

Les sujets nous ont cependant fait part de leur difficulté à faire le test. Deux problèmes majeurs se sont posés : le fait qu'A et B correspondent à une même voix et X à une voix différente les a perturbés et le problème de mémorisation des stimuli a également été important. Nous avons donc, comme pour la synthèse, mis en place un test sur la détection du changement de locuteur.

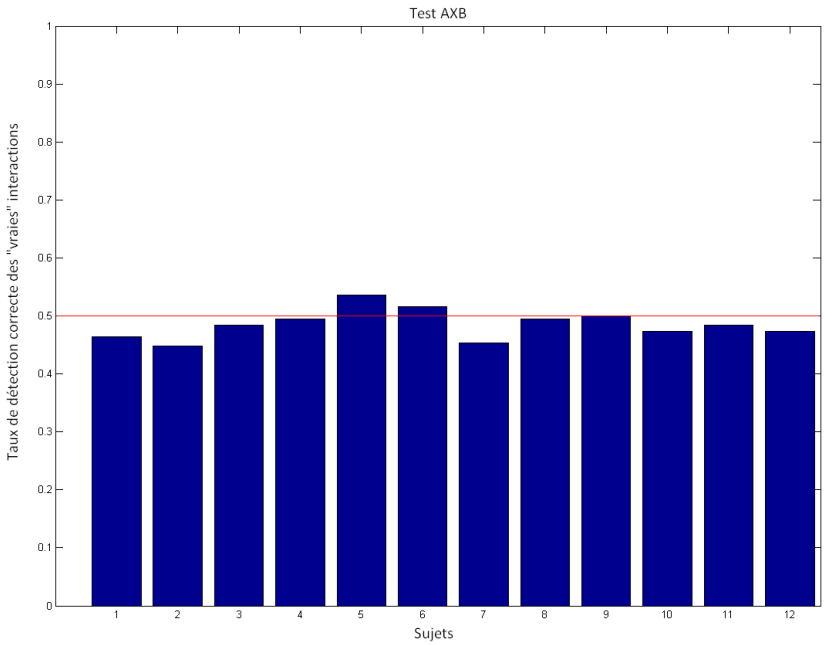


Figure 5. 6. Taux de détections correctes de « vraies vs. Fausses » interactions pendant un test AXB. Pendant ce test, on présente un stimulus du locuteur 1 prononcé pendant son interaction avec le locuteur 2 (A) puis le même stimulus prononcé par le locuteur 2 pendant son interaction avec le locuteur 1 (X) et enfin ce stimulus prononcé par le locuteur 1 pendant son interaction avec le locuteur 3 (B). On demande aux sujets testés de choisir entre A et B celui qui ressemble le plus à X. S'ils sont sensibles à la convergence leur choix devrait se porter vers A. On remarque qu'ici les sujets répondant au hasard à cause d'une charge cognitive trop importante.

5.4.1.2 Changement de locuteur

Pour faciliter la tâche pour les sujets et diminuer la charge cognitive, nous avons mis en place une tâche de changement de locuteur. Nous avons enchaîné 178 stimuli provenant des pré-tests de nos deux locuteurs étudiés ainsi que de leurs signaux d'interaction. On demande alors aux sujets de presser la barre espace s'ils détectent un changement de locuteur. On suppose que s'il y a eu une adaptation pendant l'interaction il leur sera alors plus difficile de détecter un changement de locuteur (voir la Figure 5. 7). Nous avons utilisé un ISI de 300 ms pour le test AXB et de 1000 ms de pour la détection du changement de locuteur. Nous avons volontairement augmenté la durée de l'ISI pour laisser le temps au sujet pour prendre la décision de presser ou non la barre espace.

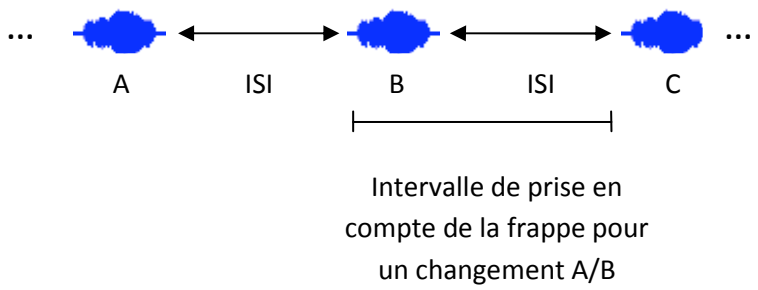


Figure 5. 7. Design utilisé pour la mise en place du test de perception sur la détection du changement de locuteur en interaction réelle. Ici l'ISI vaut 1000 ms.

Douze sujets, âgés en moyenne de 27 ans et 5 mois (quatre femmes et huit hommes), ont passé le test. Six d'entre eux connaissaient les voix de référence utilisées pour mettre en place le test et les six autres ne connaissaient pas les voix. Nous avons calculé les matrices représentant les pourcentages de fausses détections en fonction des transitions. Ces pourcentages devraient être faibles sur la diagonale ainsi qu'au coin extérieur de la matrice car cela correspond à des signaux qui proviennent soit du même locuteur dans la même condition (pré-test vs. Interaction) pour la diagonale soit des deux locuteurs en pré-test donc c'est dans ce cas que la distinction entre les signaux doit être la plus facile. La Table 5. 1 regroupe les résultats obtenus. On remarque que les sujets commettent plus d'erreurs de détection de transition avec les signaux provenant de l'interaction (33.3% d'erreurs pour les transitions loc1_loc2 puis loc2_loc1 et 18.1% d'erreurs pour les transitions loc2_loc1 puis loc1_loc2), ils ont plus de mal à distinguer de différences entre les deux voix et sont donc sensibles à l'adaptation. On observe une asymétrie dans les scores obtenus qui peut s'expliquer par une adaptation plus forte du sujet 1 vers le sujet 2 comme le montre la Figure 4. 3 (paire 14). Il est également plus difficile pour les sujets de détecter une transition entre le signal d'interaction d'un locuteur et le pré-test de son interlocuteur ou son propre pré-test (25%, 24%, 29.2% pour le premier cas et 11.9%, 14.6% et 19% pour le second cas). Dans les deux cas, s'il y a eu adaptation, les signaux d'interaction correspondent à des formes intermédiaires entre les deux pré-tests ce qui complique la tâche de détection de transition. Enfin, on observe bien des erreurs faibles sur la diagonale de la matrice, on obtient cependant des valeurs plus fortes pour les transitions entre les deux pré-tests. Nous avons donc distingué les cas des personnes connaissant les voix de celles qui ne les connaissaient pas. Les résultats sont regroupés dans la Table 5. 2 pour les personnes connaissant les voix de référence et la Table 5. 3 pour les autres et sont illustrés sur la Figure 5. 8.

	Loc1_pretest	Loc1_Loc2	Loc2_Loc1	Loc2_pretest
Loc1_pretest	4.9	11.9	25	17.7
Loc1_Loc2	6.9	4.5	33.3	6.7
Loc2_Loc1	24	18.1	7.1	19
Loc2_pretest	19	29.2	14.6	0.7

Table 5. 1. Pourcentage de fausse détection pour les 12 sujets testés. On remarque que les sujets sont sensibles à l'adaptation puisqu'ils ont plus de difficultés à détecter une transition entre les signaux d'interaction des partenaires testés. Comme prévu, les valeurs obtenues sur la diagonale sont faibles, on observe cependant des taux d'erreurs élevés pour les transitions entre les deux pré-tests.

	Loc1_pretest	Loc1_Loc2	Loc2_Loc1	Loc2_pretest
Loc1_pretest	2	4.8	19.4	8.3
Loc1_Loc2	0	3.8	28.6	6.7
Loc2_Loc1	18.8	8.3	3.8	2.4
Loc2_pretest	7.1	22.2	4.2	0.7

Table 5. 2. Pourcentage de fausse détection pour les 6 sujets testés qui connaissaient les voix de référence. Dans ce cas, les valeurs correspondant aux transitions entre les pré-tests des deux locuteurs sont cohérentes.

	Loc1_pretest	Loc1_Loc2	Loc2_Loc1	Loc2_pretest
Loc1_pretest	6.9	19	30.6	27.1
Loc1_Loc2	13.9	5.3	38.1	6.7
Loc2_Loc1	29.2	27.8	10.3	35.7
Loc2_pretest	31	36.1	25	0.7

Table 5. 3. Pourcentage de fausse détection pour les 6 sujets testés qui ne connaissaient pas les voix de référence. On remarque que le nombre d'erreurs élevé pour les transitions entre les pré-tests des locuteurs proviennent de ces sujets.

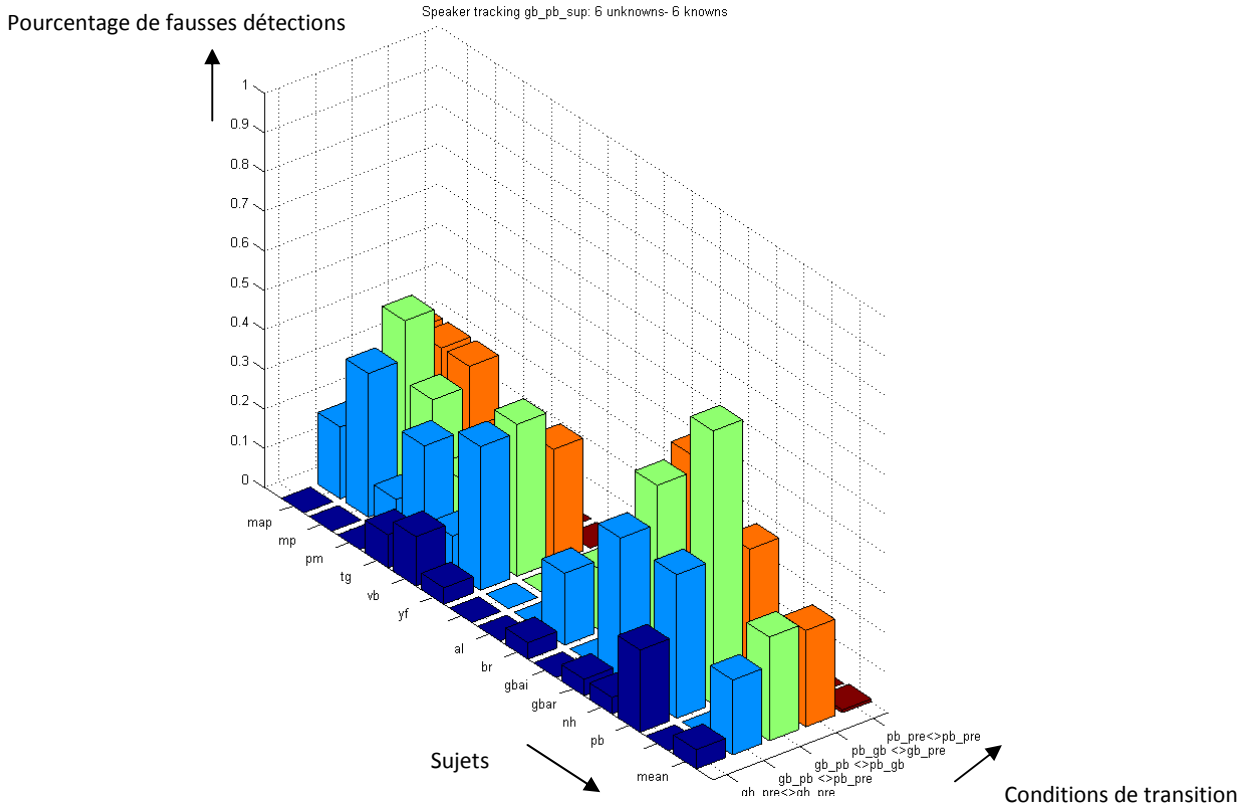


Figure 5. 8. Pourcentage de fausses détections pour différentes transitions. Les données sur le devant de la figure correspondent aux sujets qui connaissaient les voix et celles au fond de l'image correspondant aux sujets qui ne connaissaient pas les voix. De gauche à droite : nous avons d'abord les transitions entre les items prononcé par le locuteur 1 pendant son pré-test (gb_pre <> gb_pre), les transitions entre les signaux d'interaction du locuteur 1 avec les signaux prononcé pendant le pré-test du locuteur 2 (gb_pb <> pb_pre), les transitions entre les signaux d'interaction des deux locuteurs (gb_pb <> pb_gb), puis les transitions entre les signaux d'interaction du locuteur 2 avec les signaux prononcé pendant le pré-test du locuteur 1 (pb_gb <> gb_pre) et enfin celles entre les items prononcé par le locuteur 2 pendant son pré-test (pb_pre <> pb_pre). On remarque que les sujets ont plus de mal à détecter des transitions entre les signaux d'interaction (barres vertes plus hautes), cela prouve ainsi qu'ils ont sensible perceptivement au rapprochement des signaux.

On remarque que ces valeurs élevées pour la détection de transition entre les signaux provenant des pré-tests proviennent des tests faits par les personnes ne connaissant pas les voix. Cela est cohérent puisque les personnes qui connaissent les voix possèdent déjà des modèles internes de celles-ci qui vont leur permettre de faire plus facilement la distinction entre les deux voix. Nous avions pris soin de faire écouter aux sujets testés 10 stimuli provenant de chaque voix afin de les familiariser avec celles-ci. Il faut envisager d'en faire écouter un plus grand nombre.

5.4.2 En synthèse

5.4.2.1 Test AXB

Nous avons d'abord mis en place un test AXB semblable à celui utilisé par Pardo (*Pardo, 2006*) pendant lequel le X correspondait à un stimulus de notre sujet 2 synthétisé à 20% vers le sujet 1 alors que le A et le B correspondaient au même stimulus mais prononcé par le sujet 1 et synthétiser respectivement à 0 et 20% vers le sujet 2. La Figure 5. 9 illustre le déroulement du test.

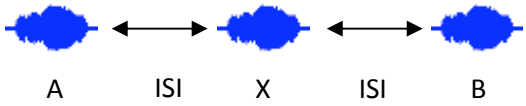


Figure 5. 9. Test AXB utilisé pour tester la perception de la convergence phonétique. A et B correspondent à des signaux obtenus en synthétisant à 0 et 20% les signaux de notre locuteur 1 vers le locuteur 2 et X correspond à un signal de synthèse créé à partir du locuteur 2 et adapté à 20% vers le locuteur 1. Les rôles des locuteurs 1 et 2 ont été inversés. Ici l'ISI vaut 300 ms.

Nous avons utilisé un intervalle inter stimuli de 300 ms pour que les sujets puissent se souvenir des trois mots prononcés et nous avons pris soin d'inverser A et B de manière équiprobable afin d'éviter tout biais dû à la mémoire à court terme (les sujets vont avoir tendance à choisir le dernier choix soit B s'ils ne se souviennent pas de A). Les rôles des sujets 1 et 2 ont également été inversés.

La consigne donnée aux sujets était de cliquer sur A s'ils jugeaient que le signal X ressemblait davantage à A et réciproquement pour B. 14 sujets ont été testés dont 4 femmes et 10 hommes âgés en moyenne de 27 ans et 10 mois. Six sujets ne connaissaient pas les voix utilisées pour créer les stimuli et les huit autres les connaissaient. Pour choisir ces voix, nous avons sélectionné une paire de sujets pour laquelle on obtenait une convergence phonétique de 32% pour le sujet de référence et de 13% pour le sujet testé afin de pouvoir comparer nos tests de perception en synthèse et en réelle interaction. La Figure 5. 10 nous donne les résultats obtenus. On obtient en moyenne une préférence de 57% pour le signal correspondant à la convergence à 20% ce qui reste très proche du hasard pour un test à deux choix possibles. Les préférences des sujets pour la convergence oscillent de 46% (correspondant à deux sujets qui ne connaissaient pas les voix et une qui les connaissait) et 65% (pour deux sujets qui connaissaient les deux voix).

Tous les sujets testés nous ont fait part de la difficulté du test dû à une charge cognitive trop importante. Nous avons donc mis en place un deuxième test de perception en synthèse basé sur la détection de changement de locuteur (Lelong and Bailly, 2012). Ce test nous permettra également d'analyser la qualité de notre synthèse adaptative.

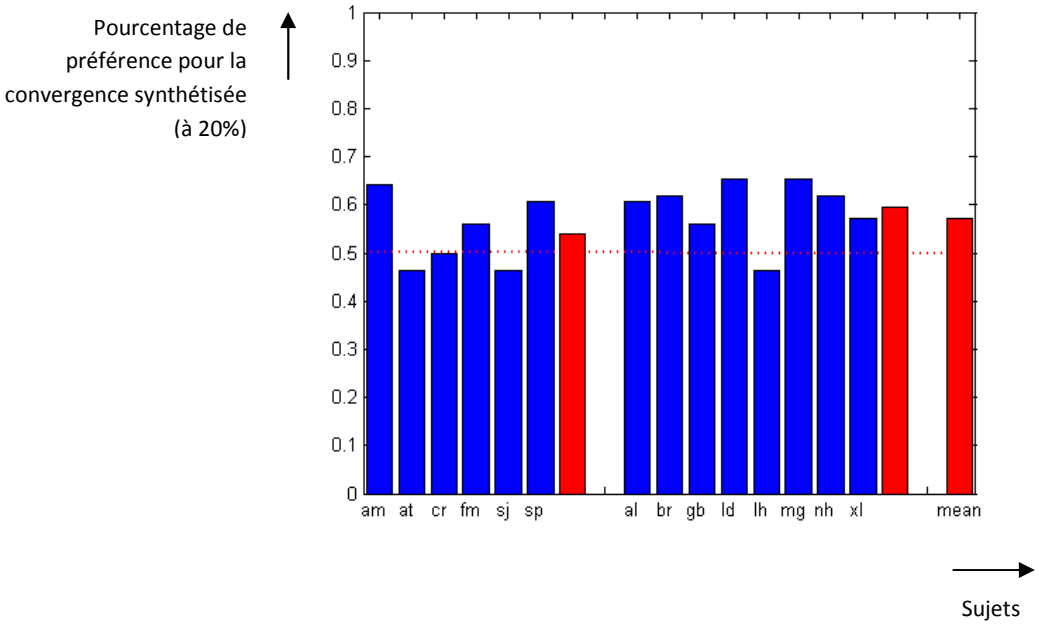


Figure 5. 10. Pourcentage de préférence pour le signal correspondant à la convergence synthétisée à 20%. On différencie les sujets qui ne connaissaient pas les voix (6 premier sujets), la barre rouge suivante correspondant à la moyenne. Les huit autres sujets connaissaient les voix utilisées pour le test, la barre rouge suivant correspondant également à la moyenne de ces huit sujets. La dernière barre rouge correspond à la moyenne sur l'ensemble des sujets. On remarque que le résultat obtenu (57%) reste très proche du seuil correspondant au hasard.

5.4.2.2 Changement de locuteur

Nous avons utilisé les mêmes stimuli que précédemment soit des stimuli de synthèse de 0 et 20% de convergence du sujet 1 vers le sujet 2 et réciproquement. Nous avons enchaîné les différents stimuli (178 au total) avec un intervalle inter-stimuli de 1 seconde pour laisser le temps aux sujets de réagir et de garantir l'attribution correcte de la décision de changement de locuteur en ligne. Nous avons pour ceci demandé aux sujets de presser la barre d'espace dès qu'ils avaient l'impression qu'il y avait eu un changement de locuteur entre le stimulus précédent et le stimulus courant. L'intervalle de validation de la frappe courre du début de la phonation du stimulus courant au début de la phonation du stimulus suivant (cf Figure 5. 7).

L'ordre des stimuli a été choisi de manière à avoir le même nombre de transitions entre nos différentes conditions (i.e. sujet 1 resynthétisé à 0%, sujet 1 synthétisé à 20% vers le sujet 2, sujet 2 synthétisé à 20% vers le sujet 1 et sujet 2 resynthétisé à 0%). Avant chaque phase de test, les sujets ont écouté 10 stimuli resynthétisé à 0% des deux locuteurs pour se familiariser avec les voix utilisées. On rappelle ici que tous les stimuli d'un même mot ont la même durée moyenne des versions pré-test de lecture de chaque locuteur.

Le test proprement dit est précédé d'une phase de familiarisation avec les voix des deux locuteurs, où deux blocs de dix stimuli du pré-test de chaque interlocuteur sont joués.

Aucune stratégie de suivi n'a été suggérée. Nous avons cependant demandé aux sujets de ne pas « corriger » une erreur de détection de changement a posteriori: la décision concerne bien la perception du changement de locuteur entre deux stimuli contigus (*speaker switching*) et non pas le

suivi de locuteur (*speaker tracking*). Il serait cependant intéressant de formaliser plus avant la stratégie de suivi en suggérant aux sujets de positionner mentalement les locuteurs dans un espace gauche-droite et de dissocier positionnement mental et détection ponctuelle.

13 sujets ont été testés dont 3 femmes et 10 hommes. Six sujets ne connaissaient pas les voix utilisées pour mettre en place le test. Nous avons remarqué qu'en majorité les sujets ont plus de mal à détecter les transitions entre les signaux qui ont été adaptés synthétiquement (locuteur 1 à 20% vers le locuteur 2 et réciproquement). Ils commettent également plus d'erreurs entre les signaux adaptés et le pré-test correspondant à l'adaptation (i.e. par exemple pour une transition entre le locuteur 1 qui a été adapté à 20% vers le locuteur 2 et le signal correspondant au pré-test du locuteur 2). Les signaux s'étant rapprochés (on a ici une convergence unidirectionnelle), il est plus difficile pour les sujets de les distinguer. La Figure 5. 11 illustre les résultats obtenus, ils sont également résumés dans les Table 5. 4 pour l'ensemble des sujets, Table 5. 5 pour les sujets qui connaissaient les voix et Table 5. 6 pour les sujets qui ne connaissaient pas les voix. Il est intéressant de remarquer que les deux sujets qui ont commis le plus d'erreur (sujets « at » et « cr ») sont ceux qui ont considéré que le test de détection de changement de locuteur était plus difficile que le test AXB. Tous les autres sujets ont eu le sentiment opposé. On remarque que les résultats, obtenus entre le test mis en place avec les signaux originaux (voir §5.4.1.2) et celui qui utilise les signaux de synthèse, sont comparables, ce qui valide la qualité de notre synthèse. De plus, sept des sujets testés n'ont pas détecté qu'il s'agissait de signaux de synthèse.

	Loc1_pretest	Loc1_Loc2	Loc2_Loc1	Loc2_pretest
Loc1_pretest	9.9	8.8	26.9	28.8
Loc1_Loc2	11.5	13.6	48.4	32.3
Loc2_Loc1	25	44.9	8.3	7.7
Loc2_pretest	19.8	35.9	7.7	8.7

Table 5. 4. Pourcentage de fausse détection pour les 13 sujets testés. On remarque que les sujets sont sensibles à l'adaptation puisqu'ils ont plus de difficultés à détecter une transition entre les signaux d'interaction des partenaires testés. Comme prévu, les valeurs obtenues sur la diagonale sont faibles, on observe cependant des taux d'erreurs élevés pour les transitions entre les deux pré-tests.

	Loc1_pretest	Loc1_Loc2	Loc2_Loc1	Loc2_pretest
Loc1_pretest	3.0	2.0	14.3	17.9
Loc1_Loc2	4.8	9.1	40.8	14.3
Loc2_Loc1	17.9	35.7	6.6	2.0
Loc2_pretest	10.2	19.0	1.8	7.1

Table 5. 5. Pourcentage de fausse détection pour les 7 sujets testés qui connaissaient les voix de référence. Dans ce cas, les valeurs correspondant aux transitions entre les pré-tests des deux locuteurs sont cohérentes.

	Loc1_pretest	Loc1_Loc2	Loc2_Loc1	Loc2_pretest
Loc1_pretest	18.1	16.7	41.7	41.7
Loc1_Loc2	19.4	18.9	10.3	53.3
Loc2_Loc1	33.3	55.6	10.3	14.3
Loc2_pretest	31.0	55.6	14.6	10.4

Table 5. 6. Pourcentage de fausse détection pour les 6 sujets testés qui ne connaissaient pas les voix de référence. On remarque que le nombre d'erreurs élevé pour les transitions entre les pré-tests des locuteurs proviennent de ces sujets. Ils ont également plus de mal à distinguer les voix même quand elles proviennent de la même personne et de la même condition.

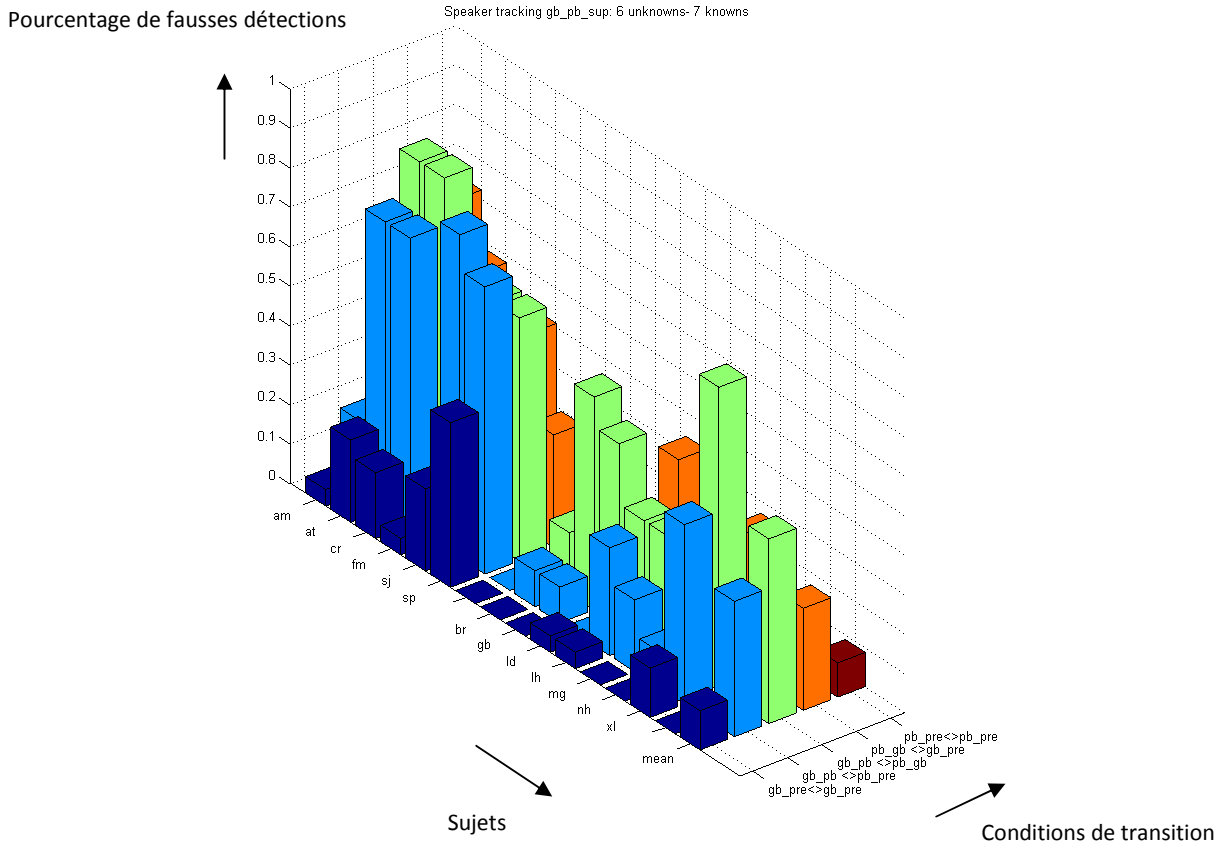


Figure 5. 11. Pourcentage de fausses détections pour différentes transitions. Les données sur le devant de la figure correspondent aux sujets qui connaissaient les voix et celles au fond de l'image correspondant aux sujets qui ne connaissaient pas les voix. De gauche à droite : nous avons d'abord les transitions entre les items prononcé par le locuteur 1 pendant son pré-test (gb_pre <> gb_pre), les transitions entre les signaux d'interaction du locuteur 1 avec les signaux prononcé pendant le pré-test du locuteur 2 (gb_pb <> pb_pre), les transitions entre les signaux d'interaction des deux locuteurs (gb_pb <> pb_gb), puis les transitions entre les signaux d'interaction du locuteur 2 avec les signaux prononcé pendant le pré-test du locuteur 1 (pb_gb <> gb_pre) et enfin celles entre les items prononcé par le locuteur 2 pendant son pré-test (pb_pre <> pb_pre). On remarque que les sujets ont plus de mal à détecter des transitions entre les signaux d'interaction (barres vertes plus hautes), cela prouve ainsi qu'ils ont sensible perceptivement au rapprochement des signaux.

5.5 Conclusions

Nous avons complété notre éventail de paradigmes expérimentaux pour caractériser la convergence phonétique avec une méthode de mesure subjective originale réalisée en ligne. La plupart des tests utilisés pour le phénomène étudié dans ce manuscrit sont des tests AXB. Nous avons donc utilisé ce type de test pour comparer nos résultats à la littérature. Nous avons effectué de nombreux tests AXB au cours de notre travail de thèse pour tenter de valider les mesures de convergence objectives que nous avons observées mais aucun test n'a donné véritablement de résultats probants dès lors que les stimuli – et notamment les styles d'élocution – étaient soit trop proches soient trop différents. Lorsque les stimuli étaient trop proches – cas où A, X et B sont tous deux des signaux d'interaction – les sujets reportent tous un problème de mémorisation des stimuli. Lorsque les stimuli étaient trop éloignés – cas où l'un des stimuli est lu et l'autre interactif – les sujets notent le style d'élocution.

Nous avons donc mis en place un nouveau type de test de perception que nous avons appelé *test de détection de changement de locuteur* (*speacker switching*). Ce test consiste à enchaîner en ordre aléatoire les stimuli provenant des diverses conditions (pré-test, interaction, etc.) et à demander aux sujets d'appuyer sur une touche du clavier lorsqu'ils détectent un changement de locuteur. Si on suppose qu'ils sont sensibles à l'adaptation des sujets en interaction, ils auront plus de mal à distinguer les transitions entre deux signaux dont les caractéristiques phonétiques convergent. Les résultats obtenus confirment cette hypothèse. Les sujets peuvent donc percevoir l'adaptation en interaction.

Notons que ce test permet d'envisager un matriçage plus complexe des conditions en intégrant des stimuli issus de répétitions, d'imitations voire d'interactions avec d'autres locuteurs que ceux impliqués dans la tâche. Suivre plus de deux locuteurs semble difficile à concevoir, à moins d'imposer explicitement une mémorisation spatiale des intervenants.

Ce chapitre présente également une première méthode de synthèse adaptative basée sur la modélisation HNM de la parole que nous avons utilisée pour calibrer et valider les tests de perception. Cette synthèse nous a permis de mettre en place des tests de perception similaires à ceux présentés précédemment et les résultats obtenus sont comparables, ce qui confirme la qualité de notre synthèse.

Conclusion et Perspectives

Rappel du contexte

Les systèmes d'interaction deviennent capables d'utiliser tous les signaux disponibles dans leur environnement pour adapter leurs réactions. Ils doivent pouvoir gérer ces signaux qui vont être impliqués dans différentes boucles de perception-action pour produire un comportement « cohérent » avec l'environnement et la tâche qu'ils doivent accomplir. De nombreuses voies de recherche complémentaires permettront de doter ces systèmes de l'intelligence sociale nécessaire. Notre approche est fondamentalement empirique: nous cherchons à construire des modèles sur des données produites en interaction homme-homme pour comprendre le fonctionnement de ces boucles d'interaction. Nos travaux ont pour but in fine de doter un agent conversationnel animé de la capacité d'adapter son comportement et plus particulièrement ses caractéristiques vocales à celles de son interlocuteur. Cette adaptation aura pour but l'amélioration de la qualité de l'interaction en transmettant, à travers l'adaptation, l'engagement de l'agent dans l'interaction avec son interlocuteur.

Démarche

La première étape de ce travail a été de mettre en place un paradigme nous permettant de récolter une quantité de données suffisamment importante pour caractériser le phénomène de convergence phonétique en interaction face-à-face. Le scénario original des « dominos verbaux », décrit dans le deuxième chapitre et dans Bailly & Lelong (2010), nous permet d'enregistrer les sujets sous différentes conditions (pré-test, interaction, répétition ambiant, post-test) tout en contrôlant le nombre d'exemplaires de chaque phonème prononcé. Ce contrôle des données a également facilité la segmentation automatique du corpus grâce à la reconnaissance de parole (avec HTK). Plusieurs types d'expériences ont été menés pour prendre en compte des critères sociaux déterminants en interaction face-à-face. Dans un premier temps, nous avons enregistré des inconnus en face-à-face médiatisé. Nous avons supposé que l'amplitude de convergence observée resterait faible puisque les sujets ne possèdent pas encore de modèles internes de leurs partenaires auxquels faire appel pour « conditionner » leur production. Dans ce cas, nous avons utilisé un premier corpus enchaînant 190 dominos (représentant environ 20 représentants de chaque cible vocalique). Nous avons testé des paires de même sexes et de sexes différents pour obtenir un critère pour recruter nos sujets (et confirmer les résultats précédents au sujet d'une convergence plus grande entre paires de femmes obtenus par Namy *et al.,* 2002). Nous avons alors remarqué (grâce à une méthode basée sur la reconnaissance de parole) que le corpus n'était pas suffisant pour l'apprentissage de modèles de Markov Cachés pour les analyses. Nous avons donc multiplié la taille du corpus par deux pour obtenir une chaîne de 350 dominos, il faudra cependant vérifier que l'augmentation de la taille du corpus n'influence pas l'amplitude de la convergence.

Dans le troisième chapitre, nous avons décrit les différentes méthodes utilisées pour mesurer de manière objective l'amplitude de la convergence phonétique. Nous avons d'abord développé une première méthode d'analyse pour mesurer le taux de convergence phonétique pour les huit voyelles orales du Français pour chaque sujet de chaque paire (Lelong et Bailly, 2011). Cette méthode est basée sur l'utilisation d'une analyse discriminante linéaire (comme l'avait fait Aubanel, 2011) sur les

coefficients MFCC extraits des pré-tests de chaque sujet (ils correspondent aux espaces phonétiques de référence de nos sujets). Nous projetons alors les données extraites de l'interaction sur le premier axe discriminant obtenu, celui-ci représentant l'espace dans lequel les données des locuteurs sont les plus séparées. Le taux de convergence correspond alors à un rapport de distances entre les données des sujets pendant le pré-test et pendant l'interaction. Cette méthode demande cependant une segmentation précise du corpus enregistré. Nous avons automatisé la segmentation en créant des modèles HMM de nos sujets à partir de leur pré-test et en utilisant ces modèles pour faire de la l'alignement de parole sur les signaux d'interaction. La segmentation a cependant été vérifiée manuellement. Nous avons donc voulu développer une méthode plus globale qui s'affranchirait de l'étape de segmentation, coûteuse en temps.

La deuxième méthode est basée sur la reconnaissance du locuteur. Nous avons utilisé pour cela la plateforme Alizée (Chartron *et al.* ,2010) développée au Laboratoire d'Informatique d'Avignon. Cette méthode utilise les modèles de mélanges gaussiens (GMM) car ce sont les modèles les plus robustes en reconnaissance du locuteur. Nous entraînons donc un GMM de chaque sujet de chaque paire en utilisant les coefficients cepstraux d'une moitié du pré-test et nous calculons le rapport de log-vraisemblance sur la deuxième moitié du pré-test et le signal d'interaction de chaque sujet en utilisant son propre modèle GMM et celui de son interlocuteur. Nous calculons alors les taux de convergence à partir des rapports de log-vraisemblance (Lelong & Bailly, 2012). Nous avons calculé les scores de corrélation entre les taux de convergence obtenus avec les deux méthodes pour valider notre seconde analyse. Nous avons trouvé des corrélations significativement élevées ($r=0.66$, $p<0.01$) pour les taux de convergence calculées par les deux méthodes. De plus, ces scores de corrélation augmentaient avec la taille de corpus. Ceci a confirmé notre conclusion au sujet de la taille du corpus. Cette nouvelle méthode de caractérisation de la convergence est très prometteuse car elle permettra de mesurer l'amplitude de la convergence phonétique en situation moins contrôlée (contenu phonétique varié) et donc plus écologique.

Les résultats obtenus avec les deux méthodes nous ont démontré que l'amplitude la convergence phonétique était plus importante pour des paires de même sexe et plus particulièrement pour des paires de femmes (peut être dû à un espace de variation plus large chez les femmes, il faudrait utiliser une normalisation des spectres ce qui est possible avec les GMMs). Nous avons observé ce phénomène dès le premier type d'expérience (i.e. entre inconnus), nous avons donc sélectionné majoritairement des paires de femmes pour la suite de nos expériences. D'autres facteurs ont également influencé l'amplitude la convergence phonétique. En effet, nous avons pu voir que le phénomène était dépendant du phonème étudié et du rôle du locuteur (i.e. le sujet testé accélère le rythme de l'interaction en fonction du taux de convergence de son partenaire, la réciproque n'étant pas vérifiée). Nous avons ensuite étudié l'impact de la « distance sociale » entre les partenaires. Nous avons remarqué que plus les partenaires étaient proches socialement (i.e. on considère que inconnus < amis < famille), plus les taux de convergence étaient élevés. En effet, plus on connaît une personne, plus on collecte d'exemplaires pour construire nos modèles internes, plus ces derniers sont riches et l'ajustement à ceux-ci est précis et rapide. Dans la même idée, nous avons étudié le lien entre la convergence phonétique et la fréquence lexicale. D'après Goldinger (1998), l'amplitude de la convergence va être plus forte pour les mots de fréquence lexicale faible. Nous collectons moins d'exemplaires des mots de fréquence lexicale faible et nous basons inconsciemment nos productions sur ces quelques exemplaires, ainsi nous avons plus tendance à imiter. Nous avons séparé notre jeu de données en deux groupes correspondant aux mots de fréquence lexicale faible et forte et calculé le

taux de convergence moyen pour chaque groupe. Nous avons obtenu le résultat attendu pour certaines paires mais cette étude mérite un approfondissement (par exemple en augmentant la taille du corpus ou en contrôlant de manière plus fine les fréquences lexicales des mots ou en utilisant de manière contrastée des mots et non-mots). Dans un dernier temps, nous avons observé l'évolution de la convergence phonétique avec le temps. Les résultats obtenus précédemment sont contradictoires. Delvaux et Soquet (2007) ainsi qu'Aubanel (2011) observent un effet cumulatif de la convergence en fonction du temps alors que Kousidis (2008) obtient une convergence immédiate des paramètres étudiés et non cumulative. De notre côté, nous avons obtenu quelques cas pour lesquels l'amplitude de la convergence augmentait avec le temps, l'évolution est cependant très lente. Il faudrait poursuivre cette analyse sur un corpus plus important car l'empan temporel ne semble pas suffisant. Une autre voie de recherches consisterait à augmenter la longueur des tours de parole en utilisant des phrases-dominos. Nous pourrions également utiliser les GMMs qui permettent de déterminer à quel moment de l'interaction il y a un changement.

Nous avons également observé l'adaptation des paramètres prosodiques (f0, durée). Pour cela, nous avons utilisé une régression linéaire des paramètres en fonction du temps pour le pré-test et pour les signaux d'interaction, nous avons alors considéré qu'il y avait une convergence des paramètres si le point d'intersection des droites de régression de chaque locuteur en interaction avait une abscisse positive et si celle était inférieure à celle obtenue avec les pré-tests. Nous obtenons quelques cas d'adaptation des paramètres prosodiques.

Enfin, nous avons voulu voir si la connaissance de la cible linguistique (comme dans le cas des répétitions qu'elles soient volontaires ou non) allait influencer l'amplitude de la convergence. On s'attend à ce que le lien entre la perception et la production pousse à imiter non volontairement le partenaire. Nous avons donc demandé aux sujets de répéter le domino précédemment prononcé par leur partenaire avant d'énoncer leur propre domino. Nous avons observé peu d'occurrence de ce phénomène. Cette étude n'a cependant été faite que sur 10 paires de sujets.

D'autres auteurs ont tenté de caractériser la convergence phonétique de manière subjective (Pardo, 2006 ; Kim, et *al.* ,2011). Le test le plus utilisé pour cela est le test AXB introduit par Goldinger (1998). Nos divers essais avec ce type de test n'ont pas été concluants, la préférence pour les signaux provenant de l'interaction avoisinant les 50% (ce qui correspond au niveau de hasard pour un test avec deux choix possibles). Les sujets nous ont fait part de leur difficulté à mémoriser les trois stimuli. Nous avons donc mis en place un paradigme original de caractérisation subjective utilisant une détection « en ligne » de changement de locuteur (voir Lelong & Bailly, ISICS 2012). Ce test suppose que le degré de convergence affecte directement le taux d'erreur de détection de transition. On remarque effectivement que les sujets ont plus de mal à distinguer les transitions entre les signaux provenant d'une interaction par rapport à ceux provenant des pré-tests correspondants. Ils sont donc sensibles à la convergence phonétique survenue pendant l'interaction. Ce test s'avère beaucoup plus intuitif et facile à appréhender par les sujets. Les taux d'erreur issus des premiers tests que nous avons effectués tant sur de la parole synthétique que naturelle sont effectivement très largement modulés par les mesures objectives de convergence. Il est également facile à mettre en place et ne nécessite pas de contraintes particulières, comme la répétition indispensable des mots pour le test AXB.

Perspectives

Les résultats présentés dans ce manuscrit sont prometteurs mais nécessitent un approfondissement. Le premier point à prendre en considération est l'augmentation de la taille du corpus. Vu la variabilité des résultats obtenus, il serait intéressant d'augmenter une nouvelle fois la taille du corpus de manière à confirmer nos résultats. Un modèle de jeu a déjà été développé à partir de mots trisyllabiques, permettant de récolter plus d'exemplaires de chaque phonème en maintenant une durée de jeu raisonnable. Pour prendre en compte la variabilité intra-locuteur, nous pourrions répéter le jeu pour chaque paire à quelques semaines d'intervalles et calculer un taux de convergence moyen sur les différentes sessions. Cela nous permettrait également de tester l'empan temporel à plus long terme. Nous avons également enregistré les signaux de post-test pour étudier le phénomène d' « after-effect », il faudra poursuivre les analyses de ces données.

Le paradigme des dominos verbaux ouvrent également de nombreuses pistes de recherche. Il nous permettrait d'étudier la convergence phonétique en utilisant une variation phonétique faible. En effet, nous pourrions utiliser des non-mots (impliquant possiblement des taux de convergence plus élevés en se basant sur la théorie des fréquences lexicales de Goldinder) pour étudier si une ambiguïté entre les deux choix possible va influencer le taux de convergence.

Les résultats obtenus entre les amis et les personnes d'une même famille soulèvent quelques questions comme par exemple à partir de quand considère-t-on qu'une voix est familière ? Dans nos expériences, les personnes qui se connaissaient depuis le moins longtemps s'étaient rencontrées 6 mois auparavant. Nous avions déjà obtenus des taux de convergence plus élevés pour cette paire. Pour tester la familiarité des voix, nous pourrions étudier l'évolution des taux de convergences de différentes paires depuis leur première rencontre jusqu'à par exemple 6 mois après, en faisant un test à chaque mois d'intervalle. Nous pourrions alors déterminer, en se basant sur les taux de convergence, à partir de quand les partenaires se considèrent comme des « amis ». Le taux de convergence pourrait donc être un bon indicateur du statut d'une relation sociale.

De plus, la question de familiarité pose également des problèmes en ce qui concerne la criminalité. En effet, si deux voix ont convergé, il sera plus difficile de distinguer entre les deux voix. A partir de quand peut-on alors considérer qu'une voix est assez familière pour la reconnaître sans erreur dans le cas d'enquête judiciaire? De plus, si une personne adapte sa voix à son interlocuteur, comment distinguer une victime de son bourreau ?

Il faut ensuite travailler sur la qualité de la synthèse obtenue. Les stimuli obtenus sont satisfaisants mais le moindre artefact pourra perturber la qualité de l'interaction homme-machine. Après avoir obtenu une qualité de synthèse adaptative correcte, il faudrait implémenter le paradigme des dominos verbaux sur la tête parlante disponible au laboratoire et mesurer les performances (temps d'accomplissement de la tâche, amplitude de la convergence) du sujet testé en fonction du degré d'adaptation multimodale de la tête parlante. Cela validera en plus l'apport de la convergence phonétique en interaction homme-machine, le but ultime étant de développer une tête parlante qui puisse s'adapter en temps réel en fonction de l'adaptation de son interlocuteur. En effet, la méthode de caractérisation basée sur la reconnaissance du locuteur permettra à la machine d'accéder en temps réel à la stratégie d'adaptation de son interlocuteur et ainsi facilitera la mise en place de sa propre stratégie d'adaptation (i.e. par exemple, la coadaptation comme pour les gestes). L'implémentation de

la convergence phonétique sur les agents conversationnels animés sera incontestablement un avantage pour les interactions homme-machine et plus particulièrement dans applications telles que des systèmes dédiés à l'apprentissage de nouvelles langues.

Pour tester notre hypothèse aux sujets des modèles internes, nous pouvons adapter le paradigme des dominos verbaux pour qu'il soit réalisable en IRM fonctionnel. Un sujet pourrait participer à l'expérience dans une condition ambiante (i.e. il interagit avec une voix préenregistrée) avec une voix qui lui est soit inconnue soit familière. On pourrait alors observer si l'activation cérébrale est différente en fonction de la condition (i.e. plus d'activation dans la condition « inconnue» qui correspond à la construction de modèles internes).

Enfin, il ne faut pas oublier que la parole est avant tout multimodale. Des premiers enregistrements des mouvements de tête ont été faits grâce au système de capture de mouvement Qualysis. L'exploitation de ce travail n'est pas présentée dans cette thèse et le lecteur est invité à se reporter aux deux articles auxquels j'ai contribué (Fagel *et al.,* 2010 & Boucher *et al.,* 2012). Des enregistrements du regard ont également été faits grâce à un oculomètre portable de marque Pertech mais sur un paradigme différent, celui-ci est présenté dans le chapitre deux. Ces enregistrements doivent encore être exploités pour enrichir les modèles d'adaptation en interaction homme-machine.

Références

Allwood, J. (2002). Bodily communication - dimensions of expression and content. Multimodality in Language and Speech Systems. B. Granström, D. House and I. Karlsson. Dordrecht, Kluwer Academic Publishers**:** 7-26.

Anderson, A., M. Bader, E. Bard, E. Boyle, G. M. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller, C. Sotillo, H. S. Thompson and R. Weinert (1991). "The HCRC Map Task Corpus." Language and Speech **34**: 351-366.

Arléo, A. (1997). Un jeu de dominos verbal: Trois p'tits chats, chapeau d'paille. Chants enfantins d'Europe. A. Arléo, A.-M. Despringre, J. Fribourg, E. Olivier and P. Panayi. Paris, L'Harmattan**:** 33-68.

Aubanel, V. (2011). Variation phonologique régionale en interaction conversationnelle. PhD Thesis. Langage et Parole; Université d'Aix-Marseille: Aix-en-Provence**:** 179 pages.

Aubanel, V., M. Cooke, J. Villegas and M. L. G. Lecumberri (2011). Conversing in the presence of a competing conversation: effects on speech production. Interspeech, Florence. Interspeech. Florence, Italy, pp. 2833-2836.

Aubanel, V. and N. Nguyen (2010). "Automatic recognition of regional phonological variation in conversational interaction." Speech Communication **52**: 577-586.

Babel, M. (2008). "The effect of talker image on phonetic convergence." Journal of Acoustical Society of America **124**: 2559.

Babel, M. E. (2009). Phonetic and social selectivity in speech accommodation. PhD Thesis. Department of Linguistics University of California: Berkeley, CA**:** 181 pages.

Bailenson, J. N. and N. Yee (2005). "Digital Chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments." Psychological Science **16**: 814-819.

Bailly, G. and A. Lelong (2010). Speech dominoes and phonetic convergence. Interspeech. Tokyo, pp. 1153-1156.

Baker, M. C. (1993). "Evidence of intraspecific vocal imitation in singing honeyeaters (Meliphagidae) and golden whistlers (Pachycephalidae)." The Condor **95**(4): 1044-1048.

Baptista, L. F. and L. Petrinovitch (1984). "Social interaction, sensitive phases and the song template hypothesis in the white-crowned sparrow." Animal Behaviour **32**: 172-181.

Bard, K. A. and C. L. Russell (1999). Evolutionary foundations of imitation: social, cognitive and developmental aspects of imitative processes in non-human primates. Imitation in Infancy. J. Nadel and G. Butterworth. Cambridge, UK, Cambridge University Press**:** 89-123.

Baron-Cohen, S. (1994). "How to build a baby that can read minds: Cognitive mechanisms in mindreading." Cahiers de Psychologie Cognitive/ Current Psychology of Cognition **13**: 513-552.

Baron-Cohen, S., D. A. Baldwin and M. Crowson (1997). "Do children with autism use the speaker's direction of gaze strategy to crack the code of language?" Child Development **68**(1): 48–57.

Baron-Cohen, S., A. Leslie and U. Frith (1985). "Does the autistic child have a "theory of mind"?" Cognition **21**: 37-46.

Bauer, J. J. and C. R. Larson (2003). "Audio-vocal responses to repetitive pitch-shift stimulation during a sustained vocalization: Improvements in methodology for the pitch-shifting technique." Journal of the Acoustic Society of America **114**(2): 1048-1054.

Bauer, J. J., J. Mittal, C. R. Larson and T. C. Hain (2006). "Vocal responses to unanticipated perturbations in voice loudness feedback: An automatic mechanism for stabilizing voice amplitude." Journal of Acoustical Society of America **119**: 2363-2371.

Bavelas, J., C. Lemery, and J. Mullet (1986). I show how you feel. Motor mimicry as a communicative act, Cognitive Psychology, vol. 50, pp. 322–329.

Bell, L., J. Gustafson and M. Heldner (2003). Prosodic adaptation in human-computer interaction. International Congress of Phonetic Sciences. Barcelona, pp. 2453-2456.

Binkofski, F., G. Buccino, S. Posse, R.J. Seitz, G. Rizzolatti, H.J. Freund, H-J (1999). A fronto-parietal circuit for object manipulation in man. Evidence from a fMRI-Study. Eur J Neurosci; 11: 3276-3286.

Binkofski, F., G**.** Buccino , K. Zilles, G.R. Fink(2004). Supramodal representation of objects and actions in the human inferior temporal and ventral premotor cortex. Cortex, 40:159-161.

Bloit J., R., X. (2008). Short-time Viterbi for online HMM decoding : evaluation on a real-time phone recognition task. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Las Vegas, NE, pp. 2121-2124.

Boersma, P., & Weenink, D. (2005). *Praat: doing phonetics by computer* (Version 4.3.01) [Computer program]. Retrieved from http://www.praat.org/.

Boucher, J.-D., U. Pattacini, A. Lelong, G. Bailly, P. F. Dominey, F. Elisei, S. Fagel and J. Ventre-Dominey (accepted). "I reach faster when I see you look: Gaze effects in human-human and human-robot face-to-face cooperation." Frontiers in neurorobotics 6, DOI:10.3389/fnbot.2012.00003.

Boula de Mareüil, P., M. Adda-Decker and C. Woehrling (2010). Antériorisation/aperture des voyelles /O/~/o/ en français du Nord et du Sud. Journées d'Etudes sur la Parole (JEP). Mons, Belgium, pp. 81-84.

Boula de Mareüil, P., B. Vieru-Dimulescu, C. Woehrling and M. Adda-Decker (2008). "Accents étrangers et régionaux en français: Caractérisation et identification." Traitement Automatique des Langues **49**(3): 135-163.

Branigan, H. P., M. J. Pickering and A. A. Cleland (2000). "Syntactic coordination in dialogue." Cognition & Emotion **75**(2): 13-25.

Branigan, H. P., M. J. Pickering, J. Pearson and J. F. McLean (2010). "Linguistic alignment between people and computers." Journal of Pragmatics.

Branigan, H. P., M. J. Pickering, J. Pearson, J. F. McLean and C. I. Nass (2003). Syntactic alignment between computers and people: the role of belief about mental states. Annual Conference of the Cognitive Science Society. Boston, MA, pp. 186-191.

Brennan, S. E. and H. H. Clark (1996). "Lexical choice and conceptual pacts in conversation." Journal of Experimental Psychology: Learning, Memory, and Cognition **22**: 1482-1493.

Buccino, G., F. Binkofski, G.R. Fink, L. Fadiga, L. Fogassi , V. Gallese, R.J. Seitz, K. Zilles, G. Rizzolatti, H.J. Freund (2001).Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. European Journal of Neuroscience 13:400-404, 2001.

Bullock, B., A. J. Toribio, V. González and A. Dalola (2006). Language dominance and performance outcomes in bilingual pronunciation. Generative Approaches to Second Language Acquisition. Somerville, MA, Cascadilla Press, pp. 9-16.

Burnett, T. A., M. B. Freedland, C. R. Larson and T. C. Hain (1998). "Voice f0 responses to manipulations in pitch feedback." Journal of the Acoustic Society of America **103**(6): 3153-3161.

Cai, S., S. S. Ghosh, F. H. Guenther and J. S. Perkell (2011). "Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing." Journal of Neurosciences **31**(45): 16483-16490.

Call, J. and M. Tomasello (2008). "Does the chimpanzee have a theory of mind? 30 years later." Trends in Cognitive Science **12**: 187-192.

Callan, D. E., J. A. Jones, A. M. Callan and R. Akahane-Yamadaa (2004). "Phonetic perceptual identification by native- and second-language speakers differentially activates brain regions involved with acoustic phonetic processing and those involved with articulatory-auditory/orosensory internal models." NeuroImage **22**: 1182-1194.

Campbell, N. (2004). Listening between the lines; a study of paralinguistic information carried by tone-of-voice. International Symposium on Tonal Aspects of Languages: Emphasis on Tone Languages. Beijing.

Chang, C. B. (2011). Systemic drift of L1 vowels in novice L2 learners. International Congress of Phonetic Sciences. Hong-Kong, pp. 428-443.

Charton, E., et al. Mistral: an open source biometric platform in 25th Symposium on Applied Computing (SAC). 2010. Switzerland.

Chistovich, L. A., G. Fant, A. d. Serpa-Leit and ao (1966a). Mimicking and perception of synthetic vowels, Part II. Stockholm - Sweden, Speech Transmission Laboratory - Department of Speech Communication and Music Acoustics - KTH**:** 1-8.

Chistovich, L. A., G. Fant, A. d. Serpa-Leit, ao and P. Tjernlund (1966b). Mimicking and perception of synthetic vowels, Part I. Stockholm - Sweden, Speech Transmission Laboratory - Department of Speech Communication and Music Acoustics - KTH**:** 1-18.

Cibelli, E. (2009). Phonetic convergence during conversational interaction in bilingual speakers. B.A. in Cognitive and Linguistic SciencesUniversity of California: Berkeley**:** 80 pages.

Clark, H. H. (1996). Using Language. Cambridge, UK, Cambridge University Press.

Content, A., C. Meunier, R. K. Kearns and U. H. Frauenfelder (2001). "Sequence detection in pseudowords in French: Where is the syllable effect?" Language and Cognitive Processes **16**(5-6): 609-636.

Coulston, R., S. Oviatt and C. Darves (2002). Amplitude convergence in children's conversational speech with animated personas. ICSLP. Boulder - Colorado, pp. 2689-2692.

Coveney, A. (2001). The Sounds of Contemporary French : Articulation and Diversity. Exeter, UK, Elm Bank Publications.

Crowne, D. P. and D. Marlowe (1960). "A new scale of social desirability independent of psychopathology." Journal of Consulting Psychology **24**: 349-354.

Dehaene-Lambertz, G., C. Pallier, W. Serniclaes, L. Sprenger-Charolles, A. Jobert and S. Dehaene (2005). "Neural correlates of switching from auditory to speech perception." NeuroImage **24**: 21- 33.

De Looze, C., C. Oertel, S. Rauzy and N. Campbell (2011). Measuring dynamics of mimicry by means of prosodic cues in conversational speech. International Conference on Phonetic Sciences (ICPhS). Hong Kong, pp. 1294-1297.

Delvaux, V. and A. Soquet (2007). "The influence of ambient speech on adult speech productions through unintentional imitation." Phonetica **64**: 145-173.

DeWolfe, B. B., L. F. Baptista and L. Petrinovich (1989). "Song development and territory establishment in Nuttall's whitecrowned sparrows." Condor **91**: 397-407.

Di Pellegrino, G., L. Fadiga, L. Fogassi, V. Gallese and G. Rizzolatti (1992). "Understanding motor events: a neurophysiological study." Experimental Brain Research **91**(1): 176-180.

Edlund, J., M. Heldner and J. Hirschberg (2009). Pause and gap length in face-to-face interaction. Interspeech. Brighton, pp. 2779–2782.

Ehrsson, H.H., A. Fagergren, T. Jonsson, G. Westling, R.S. Johansson, H. Forssberg (2000).Cortical activity in precision- versus powergrip tasks: an fMRI study. J. Neurophysiol 83:528–36.

Eimas, P. D. (1985). "The perception of speech in early infancy." Scientific American **252**(1): 46-52.

Goldinger, S. D. (1998). "Echoes of echoes? An episodic theory of lexical access." Psychological Review **105**: 251-279.

Elyan, O. (1978). "Sex differences in speech style." Women Speaking **4**: 4-8.

Fagel, S., G. Bailly, F. Elisei and A. Lelong (2010). On the importance of eye gaze in a face-to-face collaborative task ACM Workshop on Affective Interaction in Natural Environments (AFFINE). Firenze, Italy, pp. 81-85.

Ferrari, P. F., V. Gallese, G. Rizzolatti and L. Fogassi (2003). "Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex." European Journal of Neuroscience **17**: 1703-1714.

Flege, J. E. (1987). "The production of “new” and “similar” phones in a foreign language: Evidence for the effect of equivalence classification." Journal of Phonetics **15**(1): 47-65.

Flege, J. E. and W. Eefting (1987). "Cross-language switching in stop consonant perception and production by Dutch speakers of English." Speech Communication **6**: 185-202.

Fogg, B. J. and C. Nass (1997). How users reciprocate to computers: an experiment that demonstrates behavior change. Conference on Human Factors in Computing Systems (CHI). Atlanta, Georgia, ACM Press, pp. 331-332.

Fowler, C. A. (1983). "Realism and unrealism: a reply." Journal of Phonetics **11**(4): 303-322.
Fowler, C. A. (1986). "An event approach to the study of speech perception from a direct-realist perspective." Journal of Phonetics **14**(1): 3-28.
Fowler, C. A. (1991). "Auditory perception is not special: We see the world, we fell the world, we hear the world." Journal of the Acoustical Society of America **89**(6): 2910-2915.
Fowler, C. A., V. Sramko, D. J. Ostry, S. A. Rowland and P. Halle (2008). "Cross language phonetic influences on the speech of French-English bilinguals." Journal of Phonetics **36**(4): 649-663.
Garnier, M., N. Henrich and H. Dubois (2010). "Influence of sound immersion and communicative interaction on the Lombard Effect." Journal of Speech, Language, and Hearing Research **53**: 588-608.
Garrod, S. and G. Doherty (1994). "Conversation, co-ordination, and convention: An empirical investigation of how groups establish linguistic conventions." Cognition & Emotion **53**: 181-215.
Gentilucci, M. (2003). "Grasp observation influences speech production." European Journal of Neuroscience **17**: 179-184.
Gentilucci, M., F. Benuzzi, M. Gangitano and S. Grimaldi (2001). "Grasp with hand and mouth: a kinematic study on healthy subjects." Journal of Neurophysiology **86**: 1885-1699.
Gentilucci, M. and P. Bernardis (2007). "Imitation during phoneme production." Neuropsychologia **45**(3): 608-615.
Gentilucci, M., P. Santunione, A. C. Roy and S. Stefanini (2004). "Execution and observation of bringing a fruit to the mouth affect syllable pronunciation." European Journal of Neuroscience **19**: 190-202.
Gibson, J. J. (1966). The senses considered as perceptual systems. Boston, MA, Houghton-Mifflin.
Giles, H. (1973). "Accent mobility: a model and some data." Anthropological Linguistics **15**: 87-105.
Giles, H. and R. Clair (1979). Language and Social Psychology. Oxford, Blackwell.
Giles, H., J. Coupland and N. Coupland (1991). Contexts of Accommodation: Developments. Cambridge, Cambridge University Press.
Giles, H., A. Mulac, J. Bradac and P. Johnson (1987). Speech accommodation theory: The first decade and beyond. Communication Yearbook. M. L. McLaughlin. London, UK, Sage Publishers. **10:** 13-48.
Goldinger, S. D. (1996). "Words and voices : Episodic traces in spoken word identification and recognition memory." Journal of Experimental Psychology: Learning, Memory, and Cognition **22**(5): 1166-1183.
Goldinger, S. D. (1998). "Echoes of echoes? An episodic theory of lexical access." Psychological Review **105**: 251-279.
Gravano, A., R. Levitan, L. Willson, S. Benus, J. Hirschberg and A. Nenkova (2011). Acoustic and prosodic correlates of social behavior. Interspeech Florence, Italy, pp. 97-100.
Greenwald, A. G., D. E. McGhee and J. L. K. Schwartz (1998). "Measuring individual differences in implicit cognition: The implicit association test." Journal of Personality and Social Psychology **74**: 1464-1480.
Gregory Jr, S. W., B. E. Green, R. M. Carrothers, K. A. Dagan and S. W. Webster (2001). "Verifying the primacy of voice fundamental frequency in social status accommodation." Language & Communication **21**: 37-60.
Gregory, S. W. (1986). "Social psychological implications of voice frequency correlations: analyzing conversation partner adaptation by computer." Social psychology quarterly **49**(3): 237-246.
Gregory, S. W. and B. R. Hoyt (1982). "Conversation partner mutual adaptation as demonstrated by Fourier series analysis." Journal of Psycholinguistic Research **11**(1): 35-46.
Gregory, S. W. and S. Webster (1996). "A nonverbal signal in voices of interview partners effectively predicts communication accommodation and social status perceptions." Journal of Personality and Social Psychology **70**: 1231-1240.
Gregory, S. W., S. Webster and G. Huang (1993). "Voice pitch and amplitude convergence as a metric of quality in dyadic interviews." Language and Communication **13**: 195-217.
Hatfield, E., J. Cacioppo R. and Rapson (1994). Emotional contagion.Cambridge University Press.
Heinks-Maldonado, T. H. and J. F. Houde (2005). "Compensatory responses to brief perturbations of speech amplitude." Acoustics Research Letters Online **6**(3): 131-137.

Heldner, M., J. Edlund, and J. Hirschberg (2010), Pitch similarity in the vicinity of backchannels KTH Speech , Music and Hearing , Stockholm , Sweden, in Proc. of Interspeech.

Hillenbrand, J., L. A. Getty, M. J. Clark and K. Wheeler (1995). "Acoustic characteristics of American English vowels." Journal of the Acoustical Society of America **97**(5): 3099-3111.

Hueber, T. (2009), Reconstitution de la parole par imagerie ultrasonore et vidéo de l'appareil vocal : vers une communication parlée silencieuse, Thèse de doctorat, Université Pierre et Marie Curie.

Johnson, K. (1997). Speech perception without speaker normalization. Talker variability in speech processing. K. Johnson and J. W. Mullennix. San Diego, CA, Academic Press**:** 145-165.

Jones, J. A. and K. G. Munhall (2000). "Perceptual calibration of F0 production: Evidence from feedback perturbation." Journal of the Acoustical Society of America **108**: 1246-1251.

Jones, J. A. and K. G. Munhall (2002). "The role of auditory feedback during phonation: Studies of Mandarin tone production." Journal of Phonetics **30**: 303-320.

Kent, R. D. (1973). "The imitation of synthetic vowels and some implications for speech memory." Phonetica **28**: 1-25.

Kim, M., W. S. Horton and A. R. Bradlow (2011). "Phonetic convergence in spontaneous conversations as a function of interlocutor language distance." Laboratory Phonology **2**: 125-156.

Kimbara, I (2008), Gesture form convergence in joint description, Journal of Nonverbal Behavior, vol. 32, pp. 123–131.

Kohler, E., C. Keysers, M. A. Umiltà, L. Fogassi, V. Gallese and G. Rizzolatti (2002). "Hearing sounds, understanding actions: action representation in mirror neurons." Science & Consciousness Review **297**: 846-848.

Kousidis, S., D. Dorran, Y. Wang, B. Vaughan, C. Cullen, D. Campbell, C. McDonnell and E. Coyle (2008). Towards measuring continuous acoustic feature convergence in unconstrained spoken dialogues. Interspeech. Brisbane, pp. 1692-1695.

Kousidis, S., D. Dorran, C. Mcdonnell and E. Coyle (2009). Time Series Analysis of Acoustic Feature Convergence in Human Dialogues. SPECOM 2009. St Petersburg, Russian Federation.

Labov, W. (2001). The anatomy of style-shifting. Style and Sociolinguistic Variation. P. Eckert and J. R. Rickford. Cambridge, UK, Cambridge University Press**:** 85-108.

Lakin, J., V. Jefferis, C. Cheng and T. Chartrand (2003). "The chameleon effect as social glue: evidence for the evolutionary significance of nonconscious mimicry." Nonverbal Behavior **27**(3): 145–162.

Lee, C., M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan (2010), Quantification of Prosodic Entrainment in Affective Spontaneous Spoken Interactions of Married Couples, in Eleventh Annual Conference of the International Speech Communication Association, no. September, pp. 793–796.

Lelong, A. and G. Bailly (2011). Study of the phenomenon of phonetic convergence thanks to speech dominoes Analysis of Verbal and Nonverbal Communication and Enactment: The Processing Issue. A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud and A. Nijholt. Berlin, Springer Verlag**:** 280-293.

Lelong, A. et G. Bailly (2011), Dominos verbaux pour étudier la convergence phonétique, *Rencontres des Jeunes Chercheurs en Parole,* Grenoble, pp 63-66.

Lelong, A., F. Elisei, G. Bailly (2010), Etude de comportements humains en face-à-face pour les agents conversationnels animés, *Journée Interaction Homme/Robot, gdr ISIS.*

Lelong, A. and G. Bailly (2012). Characterising phonetic convergence with speaker recognition techniques. The Listening Talker Workshop,Edinburgh, pp 28-31.

Lelong, A. and G. Bailly (2012). Original objective and subjective characterization of phonetic convergence. International Symposium on Imitation and Convergence in Speech. Aix-en-Provence, France.

Leslie, A. M. (1994). ToMM, ToBY, and Agency: Core architecture and domain specificity. Mapping the Mind: Domain specificity in cognition and culture. L. A. Hirschfeld and S. A. Gelman. Cambridge, Cambridge University Press**:** 119–148.

Levitan, R., J. Hirschberg, (2011), Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions, In *Interspeech*, pp 3081-3084.

Liberman, A. M. (1982). "On finding that speech is special." American Psychologist **37**(2): 148-167.
Liberman, A. M., F. S. Cooper, D. P. Schankweiler and M. Studdert-Kennedy (1967). "Perception of the speech code." Psychological Review **74**: 431-461.
Liberman, A. M. and I. G. Mattingly (1985). "The motor theory of speech perception revisited." Cognition **21**: 1-36.
Liberman, A. M. and I. G. Mattingly (1989). "A specialization for speech perception." Science **243**: 489-494.
Lindblom, B. (1987). Adaptive variability and absolute constancy in speech signals: two themes in the quest for phonetic invariance. Proceedings of the XIth International Congress of Phonetic Sciences. Tallin, Estonia, pp. 9-18.
Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H & H theory. Speech production and speech modelling. W. J. Hardcastle and A. Marchal. Dordrecht, Kluwer**:** 403-439.
Lindblom, B., S. Guion, S. Hura, S.-J. Moon and R. Willerman (1995). "Is sound change adaptive?" Revista di linguistica **7**: 5-37.
Lindblom, B. and R. Lindgren (1985). "Speaker-listener interaction and phonetic variation." PERILUS IV - Publication of the Department of Linguistics.
Linell, P. (1998). Approaching Dialogue: Talk, Interaction, and Contexts in a Dialogical Perspective
Amsterdam, Benjamins.
Lockridge, C. B. and S. E. Brennan (2002). "Addressees needs influence speakers early syntactic choices." Psychonomic Bulletin and Review **9**: 550-557.
MacDonald, E. N., R. Goldberg and K. G. Munhall (2010). "Compensations in response to real-time formant perturbations of different magnitudes." Journal of the Acoustic Society of America **127**(2): 1058-1068.
Malfait, N., P. L. Gribble and D. J. Ostry (2005). "Generalization of motor learning based on multiple field exposures and local adaptation." Journal of Neurophysiology **93**: 3327-3338.
Matarazzo, J. D. and A. N. Wiens (1967). "Interviewer influence on durations of interviewee silence." Exp. Res. Personal **2**: 56-69.
Mattar, A. A. G. and D. J. Ostry (2007). "Modifiability of generalization in dynamics learning." Journal of Neurophysiology **98**: 3321-3329.
McCartney, J. S. and R. Panneton (2005). "Four-month-olds' discrimination of voice changes in multimodal displays as a function of discrimination protocol." Infancy **7**(2): 163-182.
McFarland, D. H. (2001). "Respiratory markers of conversational interaction " Journal of Speech, Language, and Hearing Research **44**: 128-143.
Meltzoff, A. N. and M. K. Moore (1977). "Imitation of facial and manual gestures by human neonates." Science **198**: 75-78.
Meltzoff, A. N. and M. K. Moore (1983). "Newborn Infants Imitate Adult Facial Gestures." Child Development **54**: 702-709.
Ménard, L., J.-L. Schwartz and J. Aubin (2008). "Invariance and variability in the production of the height feature in French vowels." Speech Communication **50**(1): 14-28.
Mol, L., Krahmer, E., and Swerts, M. (2009). *Alignment in iconic gestures: Does it make sense?* Paper presented at the The eight international conference on auditory-visual speech processing, Norwich, United Kingdom.
Moon, S.-J. and B. Lindblom (1994). "Interaction between duration, context and speaking style in English stressed vowels." Journal of the Acoustical Society of America **96**: 40-55.
Munhall, K. G., E. N. MacDonald, S. K. Byrne and I. Johnsrude (2009). "Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate." Journal of the Acoustic Society of America **125**(1): 384-390.
Namy, L. L., L. C. Nygaard and D. Sauerteig (2002). "Gender differences in vocal accommodation: The role of perception." Journal of Language and Social Psychology **21**: 422–432.
Nass, C. and K. M. Lee (2001). "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction." Journal of Experimental Psychology **7**(3): 171-181.

Nass, C. and Y. Moon (2000). "Machines and mindlessness: social responses to computers." Journal of Social Issues **56**(1): 81-103.
Neagu, A. (1998). Analyse articulatoire du signal de parole: caractérisation des syllabes occlusive-voyelle en Français. Institut de la Communication Parlée (ICP); Institut National Polytechnique: Grenoble - France**:** 232 pages.
Nenkova, A., A. Gravano, and J. Hirschberg (2008), High frequency word entrainment in spoken dialogue, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies Short Papers - HLT '08, p. 169.
Nygaard, L. C. and J. S. Queen (2000). "Surface form typicality and asymmetries in recognition memory." Journal of Experimental Psychology: Learning, Memory & Cognition **26**: 1228-1244.
Oertel, C., F. Cummins, N. Campbell, J. Edlund, P. Wagner, (2010), D64: A corpus of richly recorded conversational interaction. *LREC*, pp 27-30.
Ohala, J. J. (1986). "Against the direct realist view of speech perception." Journal of Phonetics **14**: 75-82.
Oviatt, S., C. Darves and R. Coulston (2004). "Toward adaptive conversational interfaces: modeling speech convergence with animated personas." ACM Transactions on Computer-Human Interaction **11**: 300-328.
Oviatt, S. L. and K. Kuhn (1998). Referential features and linguistic indirection in multimodal language. International Conference on Spoken Language Processing. Sydney, pp. 2339-2342.
Pardo, J. S. (2006). "On phonetic convergence during conversational interaction." Journal of the Acoustical Association of America **119**(4): 2382–2393.
Pardo, J. S., I. Cajori Jay and R. M. Krauss (2010). "Conversational role influences speech imitation." Attention, Perception, & Psychophysics **72**: 2254-2264.
Pardo, J. S., R. Gibbons, A. Suppes and R. M. Krauss (2012). "Phonetic convergence in college roommates." Journal of Phonetics **40**(1): 190-197.
Pearson, J., J. Hu, H. P. Branigan, M. J. Pickering and C. I. Nass (2006). Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. Conference on Human Factors in Computing Systems (CHI). Montréal.
Piaget, J. P. (1962). Play, dreams, and imitation in childhood. New York, Norton.
Pickering, M., H. Branigan, A. Cleland and A. Stewart (2000). "Activation of syntactic priming during language production." Journal of Psycholinguistic Research **29**(2): 205–216.
Pickering, M. J. and S. Garrod (2004). "Toward a mechanistic psychology of dialogue." Behavioral and Brain Sciences **27**: 169-225.
Pickering, M. J. and S. Garrod (2006). "Alignment as the basis for successful communication." Research on Language and Computation **4**(2-3): 203-228.
Pierrehumbert, J. B. (2001). Exemplar dynamics : Word frequency, lenition and contrast. Frequency effects and the emergence of linguistic structure. J. Bybee and P. Hopper. Amsterdam, John Benjamins**:** 137-157.
Podos, J. (1996). "Motor constraints on vocal development in a songbird." Animal Behaviour **51**: 1061–1070.
Purcell, D. and K. Munhall (2006). "Adaptive control of vowel formant frequency: evidence from real-time formant manipulation." Journal of the Acoustic Society of America **120**(2): 966-977.
Putland, D. A., J. A. Nicholls, M. J. Noad and A. W. Goldizen (2006). "Imitating the neighbours: vocal dialect matching in a mimic–model system." Biology Letters **2**(3): 367-370.
Repp, B. H. and D. R. Williams (1985). "Categorical trends in vowel imitation: preliminary observations from a replication experiment." Speech Communication **4**: 105-120.
Repp, B. H. and D. R. Williams (1987). "Categorical tendencies in imitating self-produced isolated vowels." Speech Communication **6**: 1-14.
Richardson, D., R. Dale and K. Shockley (2008). Synchrony and swing in conversation: coordination, temporal dynamics and communication. Embodied Communication. I. Wachsmuth, M. Lenzen and G. Knoblich, Oxford University Press**:** 75-93.

Rizzolatti, G. and M. Arbib (1999). "From grasping to speech: Imitation might provide a missing link." Trends In Neurosciences **22**(4): 151-152.
Rizzolatti, G. and M. A. Arbib (1998). "Language within our grasp." Trends In Neurosciences **21**: 188-194.
Rizzolatti, G. and L. Craighero (2004). "The mirror-neuron system." Annual Review of Neuroscience **27**: 169-192.
Rizzolatti, G. and C. Sinigaglia (2006). So quel che fai. Il cervello che agisce e i neuroni specchio. Bologna, Raffaello Cortina Editore.
Sancier, M. L. and C. A. Fowler (1997). "Gestural drift in a bilingual speaker of Portuguese and English." Journal of Phonetics **25**: 421-436.
Sato, M., K. Grabski, L. Garnier, L. Granjon, J.-L. Schwartz and N. Nguyen (2011). Plasticity of auditory goals in speech production: behavioral evidence from phonetic convergence and speech imitation. International Seminar on Speech Production. Montreal.
Sato, M., K. Grabski, L. Granjon, J.-L. Schwartz and N. Nguyen (2010). Converging to a common speech code: automatic imitative and perceptuo-motor recalibration processes in speech communication. Second Neurobiology of Language Conference. San Diego, USA.
Schwartz, J.-L., A. Basirat, L. Ménard and M. Sato (2010). "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception." Journal of Neurolinguistics: in press.
Serkhane, J. E. (2005). Un bébé androïde vocalisant: Etude et modélisation des mécanismes d'exploration vocale et d'imitation orofaciale dans le développement de la parole. PhD Thesis. Institut de la Communication Parlée; Institut National Polytechnique: Grenoble**:** 278 pages.
Simpson, A. P. (2003). Possible articulatory reasons for sex-specific differences in vowel duration. International Seminar on Speech Production. Sydney, Australia, pp. 261-266.
Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. Human Communication: A unified view. E. E. D. Jr. and P. B. Denes. New York, McGraw-Hill**:** 51-66.
Street, R. L., Jr. (1984). "Speech convergence and speech evaluation in fact-finding interviews." Human Communication Research **11**: 139-169.
Stylianou, I. (1990). Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification (PhD, Signal et Image, ENST Paris, Paris).
Suzuki, N. and Y. Katagiri (2007). "Prosodic alignment in human-computer interaction." Connection Science **19**(2): 131-141.
Tajfel, H. and J. Turner (1979). An integrative theory of intergroup conflict. The Social Psychology of Intergroup Relations. W. G. Austin and S. Worchel. Monterey, CA, Brooks-Cole**:** 94-109.
Thórisson, K. (2002). Natural turn-taking needs no manual: computational theory and model from perception to action. Multimodality in language and speech systems. B. Granström, D. House and I. Karlsson. Dordrecht, The Netherlands, Kluwer Academic: 173–207.
Thorpe, W. H. (1967). Vocal imitation and antiphonal song and its implications. Proceedings of the XVI International Ornithological Congress D. W. Snow. Oxford, UK, Blackwell**:** 245-263.
Tomasello, M. (2008). Origins of Human Communication. Cambridge, MA, MIT Press**:** 393 pages.
Tomasello, M. and A. C. Kruger (1992). "Joint attention on actions: acquiring verbs in ostensive and nonostensive contexts." Journal of Child Language **19**(2): 311-333.
Tomasello, M., S. Savage-Rumbaugh and A. C. Kruger (1993). "Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees." Child Development **34**(6): 1688-1705.
Toro, J. M., S. Sinnett and S. Soto-Faraco (2005). "Speech segmentation by statistical learning depends on attention." Cognition **97**: B25-B34.
Traum, D. and J. Allen (1992). A speech acts approach to grounding in conversation. International Conference on Spoken Language Processing (ICSLP). Banff, Alberta, Canada, pp. 137-140.
Van Vugt, H. C., J. N. Bailenson, J. F. Hoorn and E. A. Konijn (2010). "Effects of facial similarity on user responses to embodied agents." ACM Transaction on Human-Computer Interaction **17**(2): 1-27.
Ward, A. and D. Litman (2007). Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. SLaTE Workshop on Speech and Language Technology in Education. Farmington, PA.

Ward, N. and S. Nakagawa (2002). Automatic user-adaptive speaking rate selection for information delivery. International Conference on Spoken Language Processing (ICSLP). Denver, Colorado, pp. 549-552.

Wilkins, D. (2003). Why pointing with the index finger is not a universal (in sociocultural and semiotic terms). Pointing: where language, culture, and cognition meet. S. Kita. Mahwah, N.J., Lawrence Erlbaum Associates**:** 171-215.

Woehrling, C. and P. B. d. Mareüil (2006). Identification of regional accents in French: perception and categorization. International Conference on Spoken Language Processing (ICSLP). Pittsburgh pp. 1511-1514.

Woehrling, C., P. B. d. Mareüil and M. Adda-Decker (2009). Linguistically-motivated automatic classification of regional French varieties. Interspeech. Brighton, pp. 2183-2186.

Young, R and M.Frye (1966). Some are laughing; some are not why? Psychological Reports, vol. 18, pp. 747–752.

Zoltan-Ford, E. (1991). "How to get people to say and type what computers can understand." International Journal of Man-Machine Studies **34**: 527-547.

Annexe

Annexe A : Corpus I

Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2
jeûner	névé	porno	notaire	schéma	mara	torchère	sherpa
vélo	logo	thermaux	maudis	rata	tabac	patère	thermie
gauchère	sherpa	diva	vassaux	bateau	taudis	minet	neigeux
palais	lady	sauna	nature	diffus	fumé	jeudi	diva
dico	caudaux	turbot	body	mérou	roussi	valais	laineux
doser	zébu	diffus	fumeux	civet	verrou	neuraux	rôder
buffet	faitout	meulard	lardu	rousseau	sauna	débord	bordure
touret	raifort	ducat	kagou	navet	vaisseau	durcir	circée
forfait	faitout	gourou	routard	sauver	vécu	sécu	culot
toujours	journée	tarder	début	cuvée	vessie	logo	gaucher
neigée	génie	buffet	ferret	sida	d'accord	cherry	ripou
niveau	vaudou	raifort	format	corsaire	cerveau	pouffer	fétu
douma	matou	matou	toupie	vaudou	douter	tumeur	meursault
touffu	futur	pivert	verjus	ténor	normaux	sauver	verrue
turbot	bossu	jury	ripou	mauvais	verrat	rubis	bitture
support	porchère	poulet	lady	raffût	fumeux	turbot	bossu
sherpa	pari	dixcors	corsaire	meuler	légat	support	portaux
ripou	pourri	cerbère	berlue	garou	roulis	taulard	larvaire
richard	charter	lubie	bijou	livet	verrou	verjus	jurat
thermie	mica	jouter	têtu	roulet	lady	ragoût	gourou
cageot	jauger	tufeau	fauché	divers	vertu	roupie	
gecko	causer	schéma	marais	tuba	bâbord		
zébu	buffet	raifort	forfait	bordeau	doser		
ferrat	rapport	faitout	toucher	zébu	butor		

Annexe B : Corpus II

Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2
rotor	tordu	béni	niveaux	zona	nager	décor	corbeaux
durée	répit	vaudou	douma	géré	réchaud	beaucoup	courroux
pilé	létaux	mammaire	merdeux	chaumer	métaux	roulet	laineux
taudis	divers	deux coups	couvert	tauder	décor	neuraux	rôder
verseau	saumur	verdi	diguer	corbeaux	beaucoup	début	butor
murger	gémeau	guépard	partout	coucher	schéma	torture	turbot
maudit	diva	toujours	journée	massore	sortir	bômé	messie
vachard	charnu	neper	perdu	tire-feu	feuler	ciné	névé
nucaux	cautère	duvet	verrat	légat	gabie	vélo	logo
tergaux	gauchère	raccord	corsaire	bitord	tordu	gauchère	sherpa
sherpa	papou	cerbère	bergère	ducat	cador	palais	lady
poulet	lady	gerçure	surjeu	dorsaux	sauna	dico	caudaux
divers	verbaux	jeudi	divers	naja	jaloux	doser	zébu
beaucoup	couture	vertu	tubaire	loubard	barbu	buffet	faitout
turbot	bobard	bertha	tabou	butor	tordu	touret	raifort
barbu	buffet	bouder	décor	duraux	roseau	forfait	faitout
ferrat	rata	corbeau	beaucoup	zona	nabi	toujours	journée
tabou	bourru	coudée	déport	bittord	torcou	neigée	génie
rubis	biler	porter	télé	courroux	roumi	niveaux	vaudou
légère	gerçure	létaux	taudis	midi	diffus	douma	matou
surfait	ferrat	diva	vatout	furie	ripou	touffu	futur
ragout	goulu	toupie	pilou	poulie	lippu	turbot	bossu
lutter	ténu	loupé	penny	puceau	saucé	support	porchère
nullard	larget	niveaux	vaudou	sécu	curie	sherpa	pari
jet d'eau	dauber	douci	ciseau	ribord	bordé	ripou	pourri

Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2	Locuteur 1	Locuteur 2
richard	charter	jouter	têtu	divers	vertu
thermie	mica	tufeau	faucher	tuba	bâbord
cageot	jauger	schéma	marais	bordeau	doser
gecko	causer	raifort	forfait	zébu	butor
zébu	buffet	faitout	toucher	torchère	sherpa
ferrat	rapport	schéma	marra	patère	thermie
porno	notaire	rata	tabac	minet	neigeux
thermaux	maudis	bateau	taudis	jeudi	diva
diva	vassaux	diffus	fumé	valais	laineux
sauna	nature	mérou	roussi	neuraux	rôder
turbot	body	civet	verrou	débord	bordure
diffus	fumeux	rousseau	sauna	durcir	circée
meulard	lardu	navet	vaisseau	sécu	culot
ducat	kagou	sauver	vécu	logo	gaucher
gourou	routard	cuvée	vessie	cherry	ripou
tarder	début	sida	d'accord	pouffer	fétu
buffet	ferret	corsaire	cerveau	tumeur	meursault
raifort	format	vaudou	douter	sauver	verrue
matou	toupie	ténor	normaux	rubis	bitture
pivert	verjus	mauvais	verrat	turbot	bossu
jury	ripou	raffût	fumeux	support	porto
poulet	lady	meuler	légat	taulard	larvaire
dix-cors	corsaire	garou	roulis	verjus	jurat
cerbère	berlue	livet	verrou	ragou	gourou
lubie	bijou	roulet	lady	roupie	

Annexe C : Calcul des coefficients MFCC

Les coefficients MFCCs sont calculés de la manière suivante :

La procédure de calcul pas à pas des MFCC est la suivante. On découpe d'abord le signal de parole en trame (ici de 25 ms) pour obtenir des portions de signaux stationnaires, il faut également que ces trames se chevauchent afin d'éviter les transitions brusques entre chaque trame.

On procède ensuite à la pré-accentuation pour donner plus d'énergie et renforcer la contribution des hautes fréquences avec un filtre passe-haut de la forme $H(z)=1-0.9z^{-1}$ puis on fenêtre le signal avec une fenêtre de Hamming pour assurer la continuité aux bords. On calcule alors la Transformée de Fourier sur chaque trame puis on filtre par un banc de filtres triangulaires répartis le long de l'échelle de Mel. On calcule ensuite le logarithme du module de l'énergie en sortie du banc de filtres et on applique la Transformée en Cosinus Discrète inverse qui joue le rôle d'une Transformée de Fourier inverse pour obtenir tous les coefficients. On ne garde ensuite que ceux qui nous intéressent – les 12 premiers dans notre cas.